

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

EXPLORATION DE LA CAPACITÉ D'UN RÉSEAU DE NEURONES
À IMITER LE JUGEMENT ET L'EXPÉRIENCE D'UN ESTIMATEUR
CHEVRONNÉ POUR L'ATTRIBUTION DU TAUX DE PRODUCTIVITÉ
D'UNE ÉQUIPE D'EXCAVATION EN INFRASTRUCTURES MUNICIPALES

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INFORMATIQUE DE GESTION

PAR
GUY JOBIN

Juillet 2008

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.01-2006). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

TABLE DES MATIÈRES

LISTE DES FIGURES	iv
LISTE DES TABLEAUX	v
LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES.....	vi
RÉSUMÉ.....	vii
INTRODUCTION.....	1
CHAPITRE I	
L'ESTIMATION DÉTAILLÉE DES COÛTS DE PROJET	6
1.1 Généralité	6
1.2 Les fondements de l'estimation détaillée	7
1.3 Les contraintes et limites de l'estimation détaillée.....	11
1.4 L'attribution du taux de productivité : principale source d'erreur	15
1.5 Conclusion.....	17
CHAPITRE II	
LES RECHERCHES ET L'ESTIMATION DES COÛTS DE PROJET	18
2.1 Généralité	18
2.2 Pourquoi les RNA?.....	18
2.3 Résumé de l'évolution des réseaux de neurones dans le domaine de l'estimation des coûts de projet	28
2.4 Conclusion.....	35
CHAPITRE III	
LES DONNÉES BRUTES.....	38
3.1 Généralité	38
3.2 La définition du problème	40
3.3 Identification et préparation des données brutes	40
3.4 Choix de l'échantillon selon les experts du domaine	44
3.5 Conclusion.....	48

CHAPITRE IV	
TEST SUR LES RNA, RÉSULTATS ET INTERPRÉTATION	50
4.1 Généralité	50
4.2 Construction et testage du modèle.....	50
4.3 Évaluation du modèle et choix des paramètres optimaux	53
4.4 Application du modèle à l'ensemble des données avec les données d'entrées révisées.....	65
4.5 Interprétation des résultats.....	67
CONCLUSION	73
APPENDICE A	
INFORMATIONS COMPLÉMENTAIRES DE LA REVUE DE LITTÉRATURE	77
BIBLIOGRAPHIE	80

LISTE DES FIGURES

Figure		Page
1.1	Exactitude des estimations selon l'AACE en fonction des classes.....	10
1.2	Facteurs influençant la productivité d'une équipe de travail	16
2.1	Case-based reasoning (CBR)	32
2.2	Algorithme génétique.....	34

LISTE DES TABLEAUX

Tableau	Page
1.1	Classe d'estimation d'AACE 17R-978
1.2	Méthodologies d'estimation des coûts des travaux9
2.1	Sommaire des critères de sélection de la méthode d'estimation des coûts24
3.1	Données sources des rapports journaliers d'inspection.....42
3.2	Les facteurs d'influence du taux de productivité selon l'estimateur no 346
4.1	Nom des fichiers des données formatées pour l'apprentissage du RNA53
4.2	Priorité des données d'entrée pour le taux de productivité d'une équipe de tranchée53
4.3	Résultats primaires du premier essai du réseau de neurones60
4.4	Tableau comparatif des résultats finaux en heures et en pourcentage d'erreur par projet68
4.5	Tableau comparatif des résultats par tronçons de tranchée pour le réseau de neurones69
5.1	Tableau comparatif des résultats finaux en heures et en pourcentage d'erreur par projet74

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AACE	Association for the Advancement of Cost Engineering
AICQ	Association des ingénieurs-conseils du Québec
API	Application programming interface
ARFF	Attribute-Relation File Format
GNU	Licence publique générale
MLP-NN	Multi-layer perceptron neural network
MTQ	Ministère des Transports du Québec
NN	Neural network
PIREM	Principal item ration estimation method
PMI	Project Management Institute
RNA	Réseau de neurones artificiels
TCL/TK	Tool Command Language / Toolkit
WIKI	Système de gestion de contenu de site Web

RÉSUMÉ

Ce travail étudie le potentiel des RNA pour l'estimation détaillée des coûts de projet dans le domaine des infrastructures municipales.

En général, pour l'entrepreneur en construction civile, l'obtention d'un contrat se joue lors des soumissions publiques. Subséquemment, l'estimation des coûts de travaux constitue la pierre angulaire de ce processus de soumission. Cette tâche s'avère vite laborieuse considérant le nombre élevé d'éléments dont il faut tenir compte, et de plus, la valeur de chacun de ces éléments est fonction de plusieurs variables difficilement contrôlables. Il a été démontré que l'attribution du taux de productivité d'une équipe de travail est la source majeure d'erreur lors de la préparation des estimations détaillées.

Nous avons bâti un modèle de prédiction du taux de productivité d'une équipe d'installation de réseaux d'aqueduc et d'égouts. Les données qui ont servi à valider empiriquement le modèle proposé émanent de projets exécutés dans la région de Laval et des Basses-Laurentides. Un historique de données est construit à partir des rapports journaliers de surveillance des travaux de 43 projets de génie urbain. Afin d'explorer la capacité des RNA à imiter le jugement et l'expérience d'un estimateur chevronné. Deux forums de discussion ont eu lieu avec trois estimateurs du domaine pour déterminer les facteurs qui influencent le taux de productivité de l'équipe d'excavation de tranchées. Ces discussions ont permis de déterminer le jeu optimal des données d'entrée des RNA.

Trois estimateurs chevronnés ont également calculé manuellement le taux de productivité à partir des plans et devis de chacun des deux projets testés. Pour ces deux projets, les résultats de prédiction des RNA sont comparés aux résultats des trois estimateurs ainsi qu'au taux de productivité réel obtenu au chantier. Les RNA obtiennent des résultats supérieurs au niveau de la précision par rapport aux résultats des estimateurs. Des recommandations sont faites pour la préparation des futurs rapports journaliers de surveillance de travaux afin de rendre les données plus accessibles aux RNA. Ainsi, d'autres recommandations sont faites pour des recherches futures qui permettraient d'introduire le processus de prédiction dans les logiciels commerciaux d'estimation détaillée du coût de projet.

Mots-clés : réseau de neurones artificiels (RNA), estimation, génie civil, prédiction

INTRODUCTION

Dans le langage des technologies de l'information (TI), le processus d'extraction de l'information à partir d'un historique de données se nomme « fouille de données ». Frawley, d'une part, et Piatetski-Shapiro, d'autre part, définissent la fouille de données comme l'extraction d'informations ou de connaissances originales, auparavant inconnues, potentiellement utiles à partir de gros volume de données (Piatetsky-Shapiro, 1991). Quant à la SAS-Institute, elle la définit comme suit : « Un processus de sélection, d'exploration, de modification et de modélisation de grandes bases de données afin de découvrir des relations entre les données jusqu'alors inconnues. » (Han, 2001)

On peut observer plusieurs applications de la fouille de données dans différents domaines (Han, 2001). Comment les entreprises de construction en génie civil peuvent-elles tirer profit de ces technologies?

En général, pour l'entrepreneur en construction civile, l'obtention d'un contrat se joue lors des soumissions publiques. Il s'agit là d'un processus bien connu par tous les entrepreneurs en génie civil du Québec (Hegazy et Moselhi, 1995 Moselhi, Hegazy et Fazio, 1991).

Subséquentement, l'estimation des coûts de travaux constitue la pierre angulaire de ce processus de soumission. Cette tâche s'avère vite laborieuse considérant le nombre élevé d'éléments dont il faut tenir compte. En outre, la valeur de chacun de ces éléments est fonction de plusieurs variables difficilement contrôlables (El-Gafy, Taha et El-Said, 2000). Ajoutons à cela la pénurie des ressources humaines qualifiées, et on se retrouve devant un problème difficile à résoudre.

Pour l'estimation des coûts de travaux, la plupart des entreprises utilisent des logiciels spécifiques ou des outils maison développés à l'aide de logiciels de bureautique (Evans, 2004). Ces applications sont plus ou moins sophistiquées et dépendent des besoins générés par l'entreprise. Ils permettent donc de reproduire mécaniquement les tâches nécessaires à la préparation d'une estimation.

À l'étape de réalisation des travaux, l'estimation du projet est utilisée pour élaborer le budget. Le suivi de ce budget est alors effectué à l'aide de logiciels de gestion, selon les différentes méthodes de management en vigueur dans l'industrie (PMBOK, 2004). Une entreprise se spécialise dans certains types d'activité du génie civil. Pour demeurer compétitive, elle doit améliorer le processus interne de ses activités afin de devenir experte dans un domaine particulier. Cette course au perfectionnement se mesure *a posteriori* par le pourcentage de profit obtenu dans l'entreprise. Existe-t-il un moyen de lire ce perfectionnement tout au long de la réalisation du projet afin que cela puisse bonifier les prochaines estimations? Oui, car toutes les réponses se trouvent dans les données générées par l'entreprise.

Par l'utilisation de logiciels d'estimation ou de gestion, l'entreprise génère, plus ou moins consciemment, un historique de données propres à ses opérations. Or, la plupart du temps, ces historiques sont souvent perdus ou alors inutilisés, stockés dans des banques de données indépendantes ou hétérogènes, et les propriétaires ne savent pas véritablement quoi en faire, car les outils de traitement ou d'analyse sont soit inexistants, soit indisponibles commercialement. Sans ce genre d'outils informatiques, le traitement des historiques de données est impossible.

Nul doute que la préparation des estimations à partir des banques de données stockées et traitées de manière à obtenir un historique valable rendrait l'ensemble des activités de l'entreprise plus efficaces. Toutefois, à l'heure actuelle, la définition des principaux paramètres de l'estimation repose en totalité sur la rigueur, le jugement et l'expérience de la ressource, en l'occurrence l'estimateur, qui effectue le travail (Uher, 1996). Les historiques de données traitées ont la capacité de pronostiquer les valeurs des projets futurs à partir des

résultats obtenus par les projets antérieurs, et assurer ainsi une augmentation sensible de la précision et de la justesse des estimations (Han, 2001).

Le sujet de la présente recherche compare les résultats obtenus par un RNA avec les résultats obtenus par trois estimateurs chevronnés. Les entrepreneurs en génie civil utilisent en majorité l'estimation détaillée pour la préparation de leur soumission (Hegazy et Moselhi, 1995). Il sera démontré que l'attribution du taux de productivité d'une équipe de travail est la source majeure d'erreur dans ce type d'estimation (Uher, 1996). Les résultats obtenus sont des taux de productivité d'une équipe d'excavation de tranchée de conduites d'aqueduc, d'égout sanitaire et d'égout pluvial. La recherche se limite à explorer la capacité des RNA à imiter le jugement et l'expérience d'un estimateur chevronné lors de l'attribution du taux de productivité d'une équipe d'excavation de tranchée pour l'installation des conduites.

La question de recherche est la suivante :

« Est-ce qu'un modèle de réseau de neurones éprouvé a une meilleure qualité de prédiction que des estimateurs chevronnés pour l'attribution du taux de productivité d'une équipe de travail lors de l'utilisation de la méthode d'estimation détaillée? »

Répondre à la question de recherche permet de vérifier si la prédiction d'un réseau de neurones peut venir en aide aux estimateurs pour l'attribution du taux de productivité d'une équipe de travail. Si tel était le cas, l'industrie aurait une solution de rechange lors de la préparation des soumissions. Les résultats obtenus avec le réseau de neurones et avec les estimateurs d'expérience seront comparés aux productivités réelles observées aux chantiers. L'objectif ultime : déterminer si la prédiction de la fouille de données peut fournir des résultats aussi ou plus précis que les estimateurs d'expérience.

Comme les estimateurs ont à produire leurs mesures dans un court laps de temps, il demeure essentiel, pour l'introduction de ce type d'algorithme dans des logiciels commerciaux, d'obtenir de bonnes performances informatiques. Dans le cadre de cette recherche, la

performance sera seulement mesurée et répertoriée afin de fournir des données de base pour des recherches futures concernant l'optimisation de la performance.

Dans l'expectative de résultats positifs, cette recherche servira d'aide à la décision lors d'attribution des différents taux de productivité des équipes de travail. Les taux de productivité déterminent la quantité totale des heures / ressources d'une soumission (Dysert, 2004). Étant donné que les estimations des coûts de projet sont souvent imprécises en raison de ces taux de productivité, une prédiction fiable aurait pour effet d'augmenter l'exactitude de ce type d'estimation. De plus, il sera possible d'introduire l'algorithme du réseau de neurones dans les logiciels commerciaux d'estimation du coût de projet. Cette recherche pourra servir de base pour d'autres études afin de fournir à l'industrie de la construction d'autres systèmes d'aide à la décision telle que l'affectation des ressources ou la détermination du taux de profit, etc.

Objectif principal de la recherche : vérifier si on peut appliquer les techniques de la fouille de données prédictive à l'estimation détaillée et, plus particulièrement, explorer la capacité des réseaux de neurones à imiter le jugement et l'expérience d'un estimateur chevronné et ainsi fournir à l'estimateur une aide à la décision durant la préparation des estimations des coûts avec la méthode d'estimation détaillée. Afin de refléter la nature compétitive des appels d'offres, il est primordial que le système d'aide à la décision ou de prédiction d'une valeur quelconque reflète les capacités de productivité des équipes de travail de l'entreprise et non celles de l'industrie en général.

Le moyen retenu pour atteindre l'objectif est de créer un historique de données à partir des rapports journaliers de chantier de plusieurs projets et de vérifier si on obtient une précision fiable de prédiction à partir d'un algorithme de prédiction. Les algorithmes testés sont le perceptron multicouche des réseaux de neurones et la régression linéaire. La différence de précision entre les résultats du processus de fouille de données et les résultats des estimateurs chevronnés est comparée afin de servir de référence pour l'interprétation des résultats de la fouille de données.

Un des buts de la recherche : vérifier les performances informatiques du processus de la fouille de données afin d'évaluer la possibilité d'intégrer ce processus à des logiciels commerciaux d'estimations détaillées des coûts de travaux.

Ce document est divisé en quatre chapitres. Le premier chapitre décrit les caractéristiques de l'estimation détaillée. Également, on y énumère les spécifications ainsi que les limites de cette méthode d'estimation. La première partie du Chapitre II répond à la question suivante : « Pourquoi le RNA en guise d'algorithme de prédiction? ». La seconde partie, elle, résume les recherches effectuées pour l'estimation des coûts de projet. Les deux derniers chapitres présentent les essais effectués sur les données brutes ainsi que les résultats obtenus. Plus précisément, le troisième décrit le travail effectué sur les données brutes à partir des discussions avec trois estimateurs chevronnés. Quant à lui, le Chapitre IV décrit tous les essais effectués avec le RNA et la régression linéaire. À la fin de ce dernier, il y a une interprétation des résultats obtenus, faite dans un premier temps dans le cadre du domaine de la construction puis, dans un second, dans le cadre du domaine de l'informatique. Ces chapitres ont été séparés dans le but de faire ressortir de façon distincte les éléments du domaine de la construction et ceux du domaine de l'informatique. Une conclusion mettant en relief la question posée en début de recherche et les objectifs fixés vérifie dans quelle mesure les résultats répondent à ces attentes.

CHAPITRE I

L'ESTIMATION DÉTAILLÉE DES COÛTS DE PROJET

1.1 Généralité

Ce chapitre décrit les principales méthodes d'estimation des coûts pour des projets de construction. Il est démontré que l'utilisation des différentes méthodes d'estimation est tributaire de l'étape du cycle de vie où se situe le projet. On verra que la précision d'une estimation dépend du niveau de définition du projet et que ce niveau de définition est représenté par des classes définies par l'AACE. Il existe d'autres organismes qui définissent différentes classes d'estimation telles que le MTQ ou le PMI. Ces dernières ne sont pas décrites dans ce chapitre, car elles ont beaucoup de similitudes avec celles de l'AACE. Cependant, les classes du MTQ sont utilisées uniquement pour les projets de voirie du Québec et les classes du PMI le sont pour tous les genres de projet. On verra également une description sommaire des différentes contraintes et limites de l'estimation détaillée afin de faire ressortir que l'attribution du taux de productivité est la source majeure d'erreur dans ce type d'estimation. Ce chapitre se termine sur une démonstration que la fluctuation du taux de productivité est fonction de plusieurs caractéristiques intrinsèques et extrinsèques à l'entreprise de construction. Enfin, un résumé de ce chapitre est présenté.

1.2 Les fondements de l'estimation détaillée

En ingénierie, l'estimation des coûts de projet est un processus essentiel à toutes les étapes du cycle de vie d'un projet de construction. Ce cycle peut être divisé en plusieurs étapes telles que la naissance d'un projet, l'étude de faisabilité, le concept préliminaire, la préparation des plans définitifs, la réalisation de projet et l'exploitation de projet (Genest, 2002).

Chacune de ces étapes est un processus en soi, qui exige en général une approbation de toutes les parties prenantes pour la mise en œuvre de l'étape suivante. Un des facteurs majeurs de cette prise de décision : le coût de réalisation du projet (Genest, 2002). Pour l'obtention du coût du projet, on doit avoir recours à une estimation des coûts. La précision et la justesse de ces estimations dépendent du niveau de détail du projet (Dysert, 2004). Or, ce niveau est fonction de l'étape du cycle de vie concernée. Donc, le choix de la méthode utilisée pour l'estimation des coûts de projet est prépondérant.

Le guide de pratique professionnel pour l'estimation des coûts en construction établi par l'AACE est la référence dans le marché de la construction (Dysert, 2004). Cette méthodologie propose une classification et une catégorisation des coûts d'estimation en fonction du degré de définition d'un projet. Afin d'accorder plus de richesse à l'analyse des coûts, d'autres caractéristiques sont considérées, comme le but de l'estimation, la méthodologie, le degré d'exactitude et l'effort de préparation.

Le degré de définition d'un projet représente la caractéristique de tête de la méthodologie puisque, de façon générale, les caractéristiques secondaires varient en fonction du degré de définition du projet. À partir de ces cinq caractéristiques de catégorisation, cinq classes d'estimation ont été établies, la classe 1 représentant le degré le plus élevé de définition et de maturité d'un projet. La progression entre les classes 1 à 5 représente un processus au cours duquel des estimations successives sont préparées jusqu'à ce qu'une estimation finale boucle le processus.

Les cinq classes et leur définition (AACE 17R-97) sont représentées au Tableau 1.1 que voici :

Tableau 1.1
Classe d'estimation D'AACE 17R-97 (Dysert, 2004)

CLASSE D'ESTIMATION D'AACE 17R-97					
Classe d'estimation	Caractéristique première	Caractéristiques secondaires			
	Degré de définition du projet	Objectif de l'estimation	Méthodologie	Degré d'exactitude attendu	Effort de préparation
Classe 5	0 % à 2 %	Triage (sélection) ou faisabilité	Stochastique ou jugement	4 à 20	1
Classe 4	1 % à 15 %	Étude de concept ou faisabilité	Principalement stochastique	3 à 12	2 à 4
Classe 3	10 % à 40 %	Budget, autorisation ou contrôle	Mixte, mais principalement stochastique	2 à 6	3 à 10
Classe 2	30 % à 70 %	Contrôle ou soumission	Principalement déterministe	1 à 3	5 à 20
Classe 1	50 % à 100 %	Vérification de l'estimation ou soumission	Déterministe	1	10 à 100

N.B. : Si l'écart pour une valeur de 1 est de +10 / -5 %, alors l'écart pour une valeur de 10 est de 100 / -50 %. Si l'index de coût de la valeur 1 représente 0,005 % des coûts d'un projet, alors l'index de coût pour une valeur de 100 est de 0,5 %.

Pour chacune de ces classes, il existe des méthodologies d'estimation des coûts des travaux (Dysert, 2004) comme elles sont énumérées dans le Tableau 1.2 que voici :

Tableau 1.2
Méthodologies d'estimation des coûts des travaux

A.5 – Méthodologies d’estimation des coûts des travaux			
Types	Classes AACE, 17R-97	Méthodes	Description sommaire ou exemples
Méthodologie conceptuelle ou ordre de grandeur	5 4 3 (parfois)	Prix unitaire global	Exemples
			▪ Centrale électrique (coût au kilowatt)
			▪ Coût d’un hôtel (coût par chambre)
			▪ hôpital (coût par lit)
		Dimension physique	▪ Stationnement (coût par case)
			▪ Un bâtiment de vocation donné (coût au mètre carré)
			▪ Pipeline (coût au mètre linéaire)
		Facteur d’utilisation	▪ Autoroute (coût au kilomètre)
			▪ $\$B = (\$A) (\text{Cap. B} / \text{Cap. A})^e$ Valeur de e (entre 0,5 @ 0,85)
		Méthode du ratio	▪ Plan chimique (coût équipement 10,000 \$K)
			▪ Prix exact : Installation équipement = X 0,05 500 \$K
			▪ Béton : X 0,11 (1,100 \$K)
Méthode paramétrique	▪ Structure X 0,11 (1,100 \$K)		
	▪ À partir des coûts totaux de projet connu et en fonction des coûts de variables indépendantes significatives, créer un algorithme pour des prédictions de projet futur (régression linéaire, réseau de neurones)		
Déterministe ou détaillée	3 (parfois) 2 1	Méthode détaillée	▪ Déterminer le calendrier des travaux
			▪ Déterminer les ressources humaines matérielles et équipements ainsi que leur taux
			▪ Déterminer la production des équipements de travail
			▪ Déterminer les frais directs et indirects
			▪ Déterminer les contingences
			▪ Faire une prise de quantités
			▪ Préparer une analyse des risques

Dans l'ouvrage de Larry R. Dysert (Dysert, 2004), chacune des différentes méthodes d'estimation utilisées dans l'industrie est définie. Comme résumé dans le tableau précédent, on y retrouve les méthodes d'estimation du prix unitaire global, des dimensions physiques, du facteur d'utilisation, du ratio paramétrique, et la méthode détaillée ou déterministe.

Ci-dessous, la Figure 1.1 illustre le pourcentage d'exactitude d'une estimation en fonction du pourcentage d'avancement de la conception ou de la classe d'estimation AACE (Dysert, 2004).

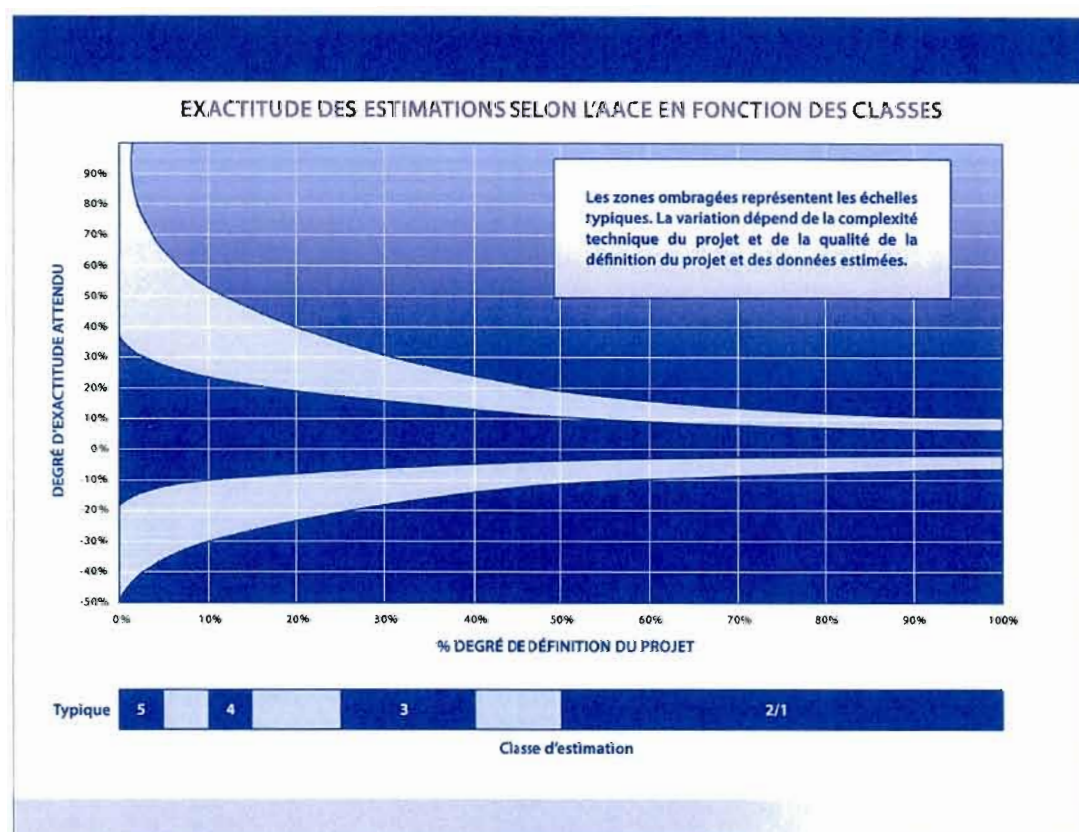


Figure 1.1 : Exactitude des estimations selon l'AACE en fonction des classes. (Traduit du Dr. Scott J. Amos, PE, de l'AACE International, 2004.)

Il sera démontré un peu plus loin dans cette étude que la plupart des travaux de recherche effectués à ce jour s'attardent en général sur la prédiction des coûts pour les estimations

paramétriques. Elles s'appliquent aux projets se retrouvant aux étapes 4 à 2 telles que décrites au Tableau 1.1. La présente recherche se concentre sur les estimations détaillées ou déterministes pour les projets avec un degré de définition de 70 @ 100 %, donc de classe 1. Plus précisément, la recherche vise à déterminer le potentiel des algorithmes de prédiction tels que les réseaux de neurones appliqués sur l'attribution du taux de productivité d'une équipe de travail dans le domaine de la construction. Afin de mesurer expérimentalement leur performance, les résultats des réseaux de neurones sont comparés aux résultats de la régression linéaire ainsi qu'aux résultats obtenus par trois estimateurs chevronnés du domaine de la construction. Les recherches en construction sur les RNA sont relativement un secteur de recherche nouveau. Cependant les résultats démontrent que les RNA sont une alternative viable à la régression linéaire parce que cette méthode élimine le besoin de trouver un historique de données qui a obligatoirement une relation mathématique entre les données (Bode, 1998, 2000 Kim, An et Kang, 2004). Cette recherche s'appuie sur cette affirmation.

1.3 Les contraintes et limites de l'estimation détaillée

Selon les classes d'estimations de l'AACE, l'estimation détaillée est une méthodologie utilisée pour les projets de classe 2 et 1 avec un degré de définition de projet de 30 @ 100 %, méthodologie utilisée pour le contrôle des coûts et les soumissions des entrepreneurs (Dysert, 2004). Pour calculer les coûts à partir de cette méthodologie, on doit déterminer le calendrier des travaux, les ressources humaines matérielles et les équipements ainsi que leur taux, la productivité des équipes de travail, les frais directs et indirects, les contingences, la prise de quantités ainsi que la préparation d'une analyse de risque. Le degré d'exactitude attendu pour une estimation détaillée pour des projets de classe 1 est de +10 @ -5 % et de +30 @ -10 % pour la classe 2, ce qui représente respectivement un écart de 15 % et de 40 % sur le coût final du projet.

Un écart de 15 @ 40 % en est un considérable dans un contexte de soumission. Si l'on considère que les marges de profit pour les entrepreneurs qui réalisent un projet de construction se situent en dessous de 5 % (Hegazy et Moselhi, 1995), au moment de la préparation des soumissions, on peut affirmer que le risque est considérable.

Peut-on améliorer le processus de l'estimation des coûts de travaux lors de la préparation des soumissions? On a vu précédemment qu'avec la méthodologie d'estimation détaillée, on doit déterminer plusieurs éléments pour l'obtention du résultat final. Quels éléments peut-on tenter d'améliorer pour obtenir des estimations plus précises? Le calendrier de travaux est souvent imposé ou tributaire des taux de productivité attribués aux équipes de travail. Il existe plusieurs méthodes précises pour la prise des quantités des ressources matérielles et leurs taux sont fonction de la connaissance du marché et leurs rabais sont tributaires de l'influence qu'exerce l'entreprise sur le marché de la construction. Quant aux ressources en équipements, elles dépendent de ceux que l'entreprise dispose ou de la capacité qu'elle a de s'en procurer d'autres. Les frais directs et indirects sont fonction respectivement de la connaissance qu'ont les dirigeants de leur entreprise et de la connaissance qu'ont les experts de ladite entreprise à réaliser le projet convoité. L'analyse de risque permet de déterminer l'incertitude à réaliser le projet soumissionné d'où découlent les frais de contingence (Dysert, 2004).

Le taux de productivité attribué à une équipe de travail découle du jugement et de l'expérience de l'estimateur et peut facilement varier d'un estimateur à l'autre. L'attribution de ce taux devient facilement un facteur de haut risque (Uher, 1996). Il est responsable du temps de réalisation pour chacune des équipes de travail. Augmenter la précision de la détermination du taux de productivité attribué à chacune d'entre elles, c'est améliorer l'ensemble de la précision de l'estimation des coûts de projet (Uher, 1996).

Pour l'estimation des coûts de travaux, la plupart des entreprises utilisent des logiciels spécifiques ou des outils maison développés à l'aide de logiciels de bureautique (Census of Construction Industries cas), lesquels sont plus ou moins sophistiqués et dépendent des besoins générés par les champs d'activité de l'entreprise.

Lors de la réalisation des travaux, l'estimation détaillée des coûts devient le budget de réalisation et son suivi est assuré à partir de logiciel comptable ou de gestion de projet, et selon les différentes méthodes de management en vigueur dans l'industrie (Dysert, 2004 PMBOK, 2004).

Il y a trois objectifs attendus pour le service de management d'un projet : les coûts à l'intérieur du budget, le calendrier des travaux en temps et la qualité telle que spécifiée. Les budgets ont un rôle clé parmi ces trois objectifs (Yu, 2005).

Une entreprise se spécialise dans certains types d'activité du génie civil. Pour demeurer compétitive, elle doit améliorer le processus interne de ses activités afin de devenir experte dans un domaine particulier. Par l'utilisation de logiciels d'estimation ou de gestion, l'entreprise génère, plus ou moins consciemment, un historique de données propres à ses opérations. Or, la plupart du temps, ces historiques sont souvent perdus ou alors inutilisés, stockés dans des banques de données indépendantes ou hétérogènes, et les propriétaires ne savent pas véritablement quoi en faire, car les outils de traitement ou d'analyse sont soit inexistants, soit indisponibles commercialement. Sans ce genre d'outil informatique, le traitement des historiques de données est inexistant

Par contre, si ces données étaient traitées de façon à obtenir un historique valable, une économie de temps majeure pourrait être obtenue en ce qui a trait à la préparation des différentes estimations des coûts de travaux. La préparation des estimations à partir des banques de données stockées et traitées de manière à obtenir un historique valable rendrait l'ensemble des activités de l'entreprise plus efficaces (Popham, 1996).

À l'heure actuelle, la définition des principaux paramètres de l'estimation repose sur le niveau de précision des méthodes de calcul utilisées et du niveau de rigueur, de jugement ou d'expérience de l'estimateur effectuant le travail (Uher, 1996). Les historiques de données ont la capacité d'informer ces estimateurs sur les résultats de projet antérieur et assureraient une augmentation considérable de la précision et de la justesse des estimations (Popham, 1996).

Toutefois, afin de déterminer une base de comparaison, la problématique centrale réside dans la difficulté de définir de façon standard le degré de similitude des mêmes activités exécutées dans le cadre de projets différents. La connaissance des projets antérieurs n'est pas toujours applicable au projet futur (Yu, 2005).

Les méthodes de travail utilisées par les entrepreneurs dans le cadre de la construction d'un projet ont un impact conséquent sur la définition des équipes de travail (à savoir affectation des ressources main-d'œuvre et des équipements) qui se voient attribuer un taux de productivité de travail (volume d'excavation à l'heure, taux de pose d'asphalte par journée, etc.). Le choix des méthodes repose principalement sur l'expérience et le jugement de l'estimateur (Hamilton et Westney, 2002). La productivité affectée à une équipe de travail découle également du jugement et de l'expérience de l'estimateur et peut facilement varier d'un estimateur à l'autre (Uher, 1996).

Il existe peu d'outils permettant l'utilisation des méthodes d'analyse systématique existantes pour mesurer l'écart survenu entre les données d'estimation des coûts de projet et les résultats du bilan d'un même projet. Cette éventualité offrirait la possibilité de valider les quantités, les méthodes de travail et les productions affectées aux équipes de travail, informations précieuses dans l'optique de la planification de projets futurs.

L'estimation détaillée est utilisée par les entrepreneurs, les gérants de projet, les promoteurs et les différents représentants du gouvernement pour l'établissement du montant total du coût de projet (Hegazy et Moselhi, 1995). L'information nécessaire provient des plans et devis et des documents d'appel d'offres. Ces estimations détaillées sont primordiales pour le suivi du budget, pour la rentabilité et pour l'atteinte des objectifs fixés par les différentes parties prenantes des projets (Yu, 2005).

Les coûts directs et indirects ainsi que les frais d'administration représentent les différents coûts associés à l'estimation détaillée (Hamilton et Westney, 2002). Les coûts directs sont composés des matériaux requis, des équipes de travail nécessaires pour la mise en place de ces matériaux et les taux correspondant à chacune de ces ressources. Les équipes de travail sont composées de la main-d'œuvre et des équipements (outil, machinerie lourde, etc.) (Dysert, 2004). Le calcul des matériaux est effectué à partir de méthodes connues et éprouvées ou une bonne précision est observée. L'affectation des équipes de travail est dictée par la nature des travaux et les équipements dont dispose le constructeur. Les coûts indirects et les frais d'administration découlent de la nature des frais directs.

L'affectation des équipes de travail et le calcul des matériaux sont effectués avec une bonne précision. L'élément majeur, où on observe une variation engendrant des imprécisions, est la production affectée à ces équipes de travail (Hancher, 1998 McCally, 1999 Uher, 1996). Augmenter la précision de la production affectée à une équipe aurait pour effet d'augmenter la précision de l'ensemble de l'estimation détaillée du coût de projet.

1.4 L'attribution du taux de productivité : principale source d'erreur

Il a été défini précédemment que l'attribution du taux de productivité était la principale source d'erreur avec la méthode d'estimation détaillée. Les variations de la productivité d'une équipe de travail de construction sont dues à plusieurs facteurs (Rowings Jr. et Sonmez, 1996). Rowings et Sonmez les regroupent en trois catégories principales. Ceux reliés à la gestion, comme l'équipe de projet, le contrôle de gestion, les méthodes et les équipements, la disponibilité du matériel et des outils, la composition des équipes, les séquences de travail, l'horaire des heures supplémentaires et la congestion; ceux reliés au projet, comme les spécifications, la conception, la grosseur des équipes, la répétition, les conditions du site, la température, l'humidité et la précipitation; ceux reliés aux tâches, comme la motivation, la morale, la fatigue, le travail syndiqué, la qualité du travail bien fait, l'absentéisme et le taux de roulement.

La compréhension de la variation du taux de productivité en fonction de la nature des travaux est difficilement maîtrisable compte tenu de la variation aléatoire des facteurs influençant l'interaction de ladite productivité avec la nature des travaux. Après analyse, les trois catégories de facteurs pourraient être graphiquement représentées selon la Figure 1.2. Les facteurs de gauche représentent les caractéristiques intrinsèques à une entreprise, soit ceux reliés à la gestion de l'entreprise et à ceux reliés aux tâches, tandis que les facteurs dans le rectangle du centre représentent les conditions particulières d'un projet. Les matériaux, la main-d'œuvre et les équipements représentent les ressources de l'entreprise pour la réalisation du projet. La productivité est la valeur recherchée. Même si l'entreprise contrôle les caractéristiques de ses ressources, celles-ci sont totalement modifiées en fonction des

caractéristiques particulières des différents projets. La variation est donc difficile à mesurer (Al-Tabtabai, Alex et Tantash, 1999 Leonard, Fazio et Moselhi, 1988 Leu, 2004).

Pour une meilleure compréhension, les différents facteurs influençant la productivité d'une équipe de travail ont été mis sous forme graphique telle que représentée à la Figure 1.2.

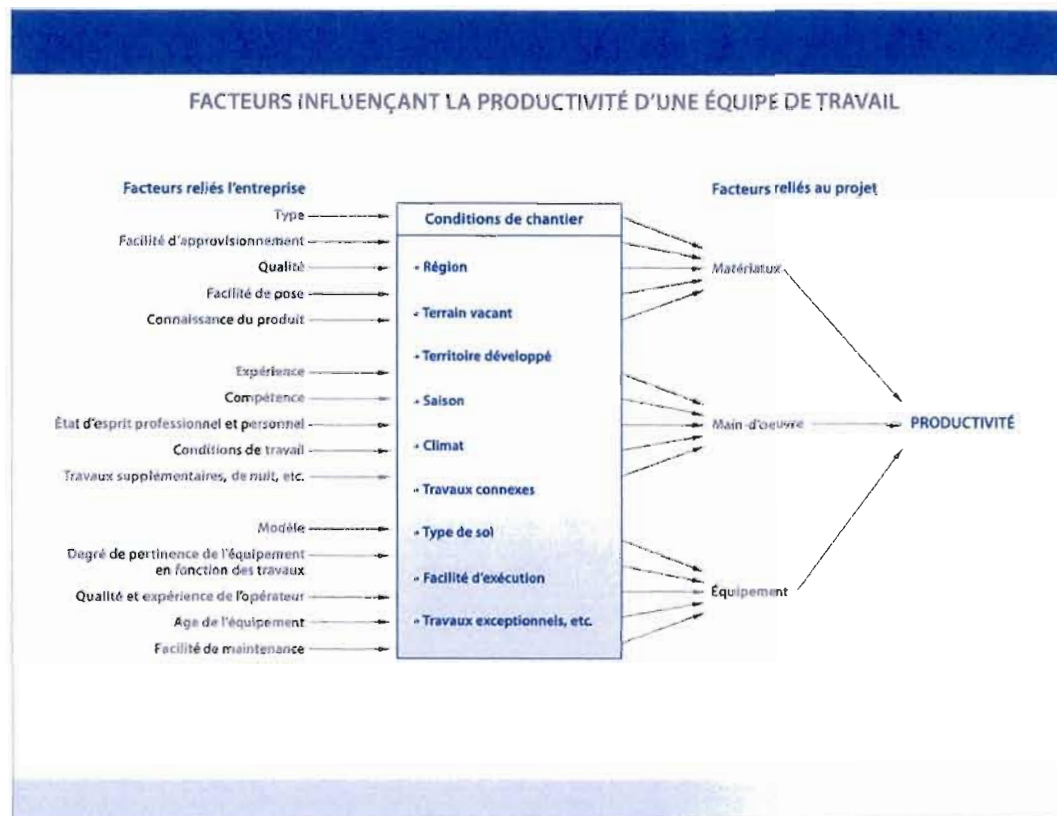


Figure 1.2 : Facteurs influençant la productivité d'une équipe de travail.

Lors de la préparation d'une soumission à partir d'une estimation détaillée, telle que nous l'avons vue précédemment, on doit attribuer un taux de productivité à une équipe de travail. La Figure 1.2 représente les différents facteurs énoncés par Rowings (Rowings Jr. et Sonmez, 1996). Ce graphique permet de voir que ceux reliés à l'entreprise, bien qu'ils fluctuent indépendamment des conditions de chantier, changent également selon les facteurs de ces différentes conditions de chantier. À titre d'exemple, un opérateur de pelle hydraulique débutant peut travailler dans des conditions normales à 70 % de la productivité d'un

opérateur bien expérimenté. S'il travaille dans des conditions de chantier extrêmes, cette variation de productivité peut atteindre des taux beaucoup plus bas. Elle est difficile à mesurer ou à prévoir, ce qui laisse présager que la variation du taux de productivité n'est pas nécessairement linéaire. Ceci est en lien avec les travaux de Singh qui conclue qu'une méthode de prédiction non linéaire est plus précise dans le domaine de l'estimation de coût de projet de construction (Singh, 1993).

1.5 Conclusion

On a vu que la méthode d'estimation des coûts était fonction du niveau de définition d'un projet. Lors de la période d'appel d'offres, les entrepreneurs préparent leur soumission afin d'obtenir la construction d'un projet. Normalement à ce stade du cycle de vie du projet, son niveau de définition est élevé et la méthode d'estimation utilisée est la méthode détaillée. L'estimation détaillée a une précision variant de plus ou moins 15 %. Le profit qu'une entreprise est en droit de s'attendre pour la construction d'un projet est de plus ou moins 5 %. Pour une entreprise de construction, le niveau de risques est donc assez élevé. On a également vu que l'attribution du taux de productivité d'une équipe de travail est la source majeure d'erreur dans ce type d'estimation. L'attribution du taux de productivité repose sur le jugement et l'expérience des estimateurs chevronnés. Trouver une méthode de prédiction fiable et précise du taux de productivité d'une équipe de travail aurait pour effet de diminuer le risque pour les entreprises de la construction.

CHAPITRE II

LES RECHERCHES ET L'ESTIMATION DES COÛTS DE PROJET

2.1 Généralité

Ce chapitre fait un survol des différentes recherches exécutées dans le domaine de l'estimation des coûts de projet. Dans un premier temps, la pertinence des RNA par rapport à d'autres méthodes est discutée. Par la suite, un résumé chronologique des différentes recherches en estimation des coûts de projet est présenté en faisant ressortir les points forts et points faibles par rapport à la prédiction des coûts de projet. Un des buts de ce résumé : faciliter la paramétrisation du réseau de neurones par l'identification de similitudes entre les recherches exécutées et la présente recherche : nombre de données dans l'historique de données, nombre de données d'entrée (variables indépendantes), précision obtenue, type de réseau de neurones utilisé, nombre de couches cachées, nombre de neurones dans la couche cachée, etc. Le chapitre se termine par un résumé des principales recherches sur lesquelles est basée la présente, incluant une discussion.

2.2 Pourquoi les RNA?

Pour identifier le système d'aide à la décision optimal pour l'attribution du taux de productivité d'une équipe de travail, plusieurs pistes sont disponibles. Entre autres, il y a l'approche probabiliste, l'analyse de risques et la prédiction numérique.

En estimation des coûts de projet, l'approche probabiliste tient compte de valeur ordinale pour le montant à prévoir des quantités et de leur coût unitaire. À titre d'exemple, au lieu d'avoir 3 500 mètres carrés de tapis à 25 \$ le mètre carré, on a plutôt une quantité de 3 200 @ 3 800 mètres carrés et un coût unitaire de 23 @ 27 \$, et ainsi de suite pour chacune des ressources de l'estimation. Puis, à partir de la méthode de Monte-Carlo, l'ensemble des coûts des ressources les plus probables est additionné pour obtenir le coût final du projet. L'ouvrage d'Uher est un bon exemple de cette méthode (Uher, 1996). Il s'attarde aux risques corporatifs et aux risques de projet dont les sources sont internes et externes (voir Appendice A.1, The General Risk Classification Method). L'article compare deux méthodes de distribution des risques dans l'estimation des coûts, soit les méthodes de distribution uniforme et celles de distribution triangulaire. Chacune utilise celle de Monte-Carlo pour distribuer le risque (voir Appendice A.2, The Proposed Probabilistic Cost Estimating Model).

Étant donné que l'attribution du taux de productivité d'une équipe de travail est tributaire de plusieurs caractéristiques intrinsèques et extrinsèques de l'entreprise et du projet, cette méthode est difficilement applicable dans ce contexte. Cependant, elle pourrait être utilisée après une autre méthode de prédiction afin d'atténuer le pourcentage d'erreur et ainsi obtenir une meilleure précision.

Akintoye *et al.* et Kartam *et al.* ont sondé par questionnaire les entrepreneurs généraux ainsi que les gestionnaires de projet de construction concernant les systèmes d'analyse de risques (Akintoye, 1997). Ces deux sondages concluent que les techniques de l'analyse de risques sont rarement utilisées dans l'industrie de la construction à cause du manque de connaissance et parce que les intervenants doutent de la capacité de ces techniques. Ils en concluent également que pour avoir un bon système de risques, on doit définir un contexte approprié ainsi que les buts et objectifs de cette analyse de risques. Compte tenu du court laps de temps pour la préparation des soumissions lors de la période d'appel d'offres (en général deux semaines) et compte tenu de l'ampleur des projets et de la variété de ceux-ci, il devient difficile pour un entrepreneur général de mettre en place un tel système (Akintoye, 1997).

Yu, Lai et Li ont développé une méthode d'estimation originale (Yu, 2005). L'objectif premier : créer une base Web intelligente pour l'estimation des coûts. Malgré que cette approche (partie Web) soit en contradiction avec un des buts de la présente recherche (créer un historique propre à une entreprise), cet article a été retenu pour son approche de calculs avec la méthode PIREM (*principal items ratio estimation method*). Elle se base sur le critère optimal de Pareto, soit que presque 80 % du temps requis pour l'estimation de coûts peut être sauvé. Le principe de base est que selon Pareto, 20 % des objets d'une estimation représentent plus de 80 % du coût total de l'estimation. Donc, pour les projets futurs, on s'attarde à 20 % des objets de l'estimation avec la méthode analogique et 80 % avec la méthode paramétrique. Cette approche pourrait devenir une solution complémentaire ou de rechange à l'approche de la prédiction numérique si celle-ci s'avérait infructueuse en partie ou en totalité.

Il existe plusieurs approches à la prédiction numérique. Les principales sont la régression (linéaire, polynomiale...) et les méthodes connexionnistes. Un avantage important des dernières par rapport à la régression est leur grande flexibilité : capacité de traiter directement des variables d'entrée non numériques (ce qui n'est pas le cas ici), des variables de distribution quelconque (pas d'hypothèse sur la distribution des données d'entrée comme la normalité) et la capacité d'estimer des fonctions quelconques, c'est-à-dire non contraintes par un modèle de départ comme en régression.

Comme nous l'avons vu au chapitre précédent, la compréhension de la variation du taux de productivité d'une équipe de travail est fonction des caractéristiques de l'entreprise jumelées aux caractéristiques du projet convoité. Compte tenu du nombre de facteurs influençant ladite productivité, on peut donc croire que cette variation n'est pas linéaire. De ce fait, les réseaux de neurones deviennent totalement appropriés. Smith et Mason sont arrivés à la conclusion que les réseaux de neurones ont un avantage significatif si les données utiles ont un rapprochement minime avec la cible estimée : ce qui est le cas (Smith et Mason, 1996). Par contre, la régression linéaire a un avantage considérable si les données utilisées sont totalement alignées avec la cible concernée : à cause des facteurs qui influencent le taux de productivité, on peut affirmer que ce n'est pas le cas.

Quels sont les avantages des réseaux de neurones artificiels sur les méthodes conventionnelles telles que la régression linéaire ou polynomiale? Un résumé des différentes conclusions des recherches analysées fournit les réponses à cette question.

Dans la recherche sur la prédiction des performances énergétiques des turbines à vent, Li *et al.* concluent en général que les réseaux de neurones performant mieux que les méthodes conventionnelles de régression (Li *et al.*, 2001). La régression requiert une fonction explicite et bien définie avant d'être en mesure d'estimer les paramètres de la méthode des moindres carrés, tandis que les réseaux de neurones dépendent plutôt des données d'entraînement et de la capacité d'apprentissage de l'algorithme.

Le modèle des variables d'entrée a été limité à la vitesse du vent à 40 mètres et à la direction du vent lu aux deux tours météo. Considérant que les données d'entrée ont été limitées à ces lectures et que plusieurs autres variables influencent l'énergie produite par une turbine à vent, il a été démontré que les réseaux de neurones dans de telles conditions performant mieux que les méthodes de régression.

La production d'énergie éolienne peut être affectée par d'autres facteurs tels que la densité de l'air, le profil vertical du vent, la saison et l'heure de la journée. Sous l'influence compliquée du nombre de facteurs de même type, le choix d'une fonction appropriée pour un modèle de régression s'avère extrêmement difficile et confère donc au réseau de neurones un avantage accru (Li *et al.*, 2001).

Ils (Li *et al.*, 2001) ont identifié que des relations évidentes ont été observées entre les modèles de régression et les réseaux de neurones quand les fonctions d'activation non linéaires sigmoïdales sont réduites à des fonctions linéaires. Un réseau de neurones général de premier ordre sera équivalent à un modèle de régression linéaire multivariable, et un réseau de neurones d'ordre supérieur sera équivalent à un modèle de régression polynomiale multivariable. Donc, la fonction d'activation non linéaire est un facteur important dans un réseau de neurones pour la résolution de problèmes complexes de fonctions d'approximation. Cette observation laisse présager que si les données d'entrée n'ont pas de lien linéaire, les

réseaux de neurones seront plus performants, ce qui rejoint les observations de Bode (Bode, 2000 Li *et al.*, 2001).

Pour le modèle de régression, la relation obtenue entre l'entrée et la sortie (valeur prédite) est une ligne droite reflétant les caractéristiques de dépendance de la fonction du modèle. Cependant, le réseau de neurones réalise la relation d'entrée-sortie à travers l'apprentissage. Avec les RNA, la relation entre les données d'entrée et les résultats obtenus sont difficilement perceptibles, contrairement à la régression linéaire.

Quand les données d'apprentissage sont sélectionnées de façon équivalente pour des vitesses de vents différentes, le réseau de neurones obtient de meilleures performances autant pour des vitesses de vent léger que pour des vitesses de vent élevé, même pour des données de test issues de mois ou d'années différentes (Li *et al.*, 2001).

Les chercheurs (Li *et al.*, 2001) ont comparé, pour la régression polynomiale, des modèles du premier jusqu'au quatrième ordre. Il est observé que la polynomiale de troisième ordre du modèle de régression fournit les meilleurs résultats de prédiction et qu'autant le plus bas que le plus élevé des ordres dégénèrent les résultats. Cela indique que la polynomiale de l'ordre inférieur n'est pas approprié pour cette application; et la polynomiale d'ordre supérieur peut introduire du bruit à une vitesse de vent élevée, même si les paramètres pour des objets d'ordre supérieur sont très petits. Le réseau de neurones est configuré avec trois couches : entrées, cachées et sorties. Le nombre de nœuds pour chacune des couches est respectivement 4, 8 et 1 nœuds. Le réseau de neurones ainsi formé performe mieux que la régression polynomiale de troisième ordre.

Smith *et al.* concluent donc qu'un réseau de neurones artificiel peut être un substitut attrayant pour la régression si l'étape du paramétrage du modèle de la régression ne peut être complétée avec succès (Smith et Mason, 1996). En d'autres mots, si les données de coût ne permettent pas l'ajustement d'un modèle de régression ou ne permettent pas à l'analyste de discerner une relation entre les coûts d'estimation de références et les coûts de projet envisagés. Le problème du paramétrage du modèle de régression devient plus complexe au

fur et à mesure que la dimension de la variable indépendante grossit, tandis que le réseau de neurones allège cette problématique. Cependant, il y a le danger possible de choisir un modèle neural trop limité, spécialement lorsque de petits échantillons sont utilisés pour l'entraînement. Bien que les performances des réseaux de neurones soient meilleures, les chercheurs concluent que l'on doit être prudent avant l'utilisation d'un réseau de neurones (Smith et Mason, 1996).

Ils mettent en garde l'utilisateur de réseaux de neurones contre certains facteurs autres que la performance de prédiction des modèles retenus. Les facteurs à considérer avant de les choisir sont la crédibilité, les questions tactiques et la capacité de reproduction des résultats (Smith et Mason, 1996).

En ce qui a trait à la crédibilité, les chercheurs énoncent qu'avec un modèle de régression, il est possible d'expliquer à la direction (ou notamment à un client) comment les résultats ont été obtenus, ce qui est impossible avec les réseaux de neurones. La prédiction de ces algorithmes est un acte de foi. Décrire l'obtention des résultats avec un réseau de neurones est comme expliquer le succès d'un joueur de tennis en disséquant son cerveau (Smith et Mason, 1996).

Concernant les questions tactiques, l'approche du réseau de neurones atténue toutes les difficultés associées aux activités d'avant-projet lorsqu'une méthode d'estimation paramétrique statistique est utilisée, sans pour autant en créer de nouvelles. Avec la régression et avant l'estimation des coûts de projet, l'analyste est confronté à choisir des coûts conducteurs, contrairement au réseau de neurones.

Pour la possibilité de reproduire les résultats, la formation d'un réseau de neurones est une procédure algorithmique et les résultats peuvent très certainement être reproduits; il faut toutefois utiliser un paramétrage identique, les mêmes poids de départ, les mêmes données de formation et la même méthode déterministe de présenter les données durant l'entraînement. Toutefois, si seulement un de ces paramètres est altéré, les résultats du réseau de neurones seront très certainement différents de l'original. Cette différence peut s'avérer mineure ou,

dans certains cas, majeure. Cela est un des aspects de la construction et de la validation du réseau de neurones. De plus, produire des modèles de réseaux de neurones optimaux demande l'identification de bonnes combinaisons architecturales, des méthodes d'entraînement performantes et des critères d'arrêt appropriés (Smith et Mason, 1996).

Pour terminer, Smith *et al.* concluent que les réseaux de neurones seront de plus en plus utilisés comme substitut à la régression par les utilisateurs de la méthode d'estimation paramétrique, puisque les analystes se rendront compte que, dans des situations particulières, les réseaux de neurones fournissent une précision d'estimation supérieure. Les réseaux de neurones seront considérés comme des solutions de rechange viables à la régression s'il y a dans les données de départ un comportement de coût sous-jacent, des données non linéaires ou des discontinuités fonctionnelles significatives, surtout pour des ensembles de données composées d'un grand nombre de variables indépendantes.

Bode résume les résultats obtenus dans la recherche de 2000 dans le Tableau 2.1, Sommaire des critères de sélection de la méthode d'estimation des coûts (Bode, 2000) :

Tableau 2.1
Sommaire des critères de sélection de la
méthode d'estimation des coûts

Utilisez le réseau neural lorsque ...	Utilisez les méthodes paramétriques lorsque...	Utilisez l'estimation des coûts détaillés lorsque...
... vous avez à votre actif plusieurs cas similaires	... vous avez à votre actif plusieurs cas similaires	... vous connaissez le nombre exact des heures et la quantité de matériel nécessaire
... vous êtes relativement certain des attributs ayant un effet sur le coût	... vous savez précisément quels attributs ont un effet sur le coût	... vous savez précisément quels attributs ont un effet sur le coût
... il y a peu de coûts conducteurs	... il y a peu de coûts conducteurs	... il y a plusieurs coûts conducteurs
... Vous ne savez pas de quelle manière les conducteurs influencent les coûts	... vous êtes relativement certain de la manière dont les conducteurs influencent les coûts	... vous savez exactement de quelle manière les conducteurs influencent les coûts

L'estimation paramétrique utilise la régression linéaire ou polynomiale et, selon Bode, l'une des difficultés majeures des méthodes d'estimation paramétriques (régression) est que l'estimateur doit connaître *a priori* une bonne partie de la fonction des coûts (Bode, 2000).

En contraste, les réseaux de neurones sont des estimateurs non paramétriques qui essaient d'ajuster les courbes des données sans avoir au préalable de fonctions prédéterminées. Les réseaux de neurones sont donc en mesure de détecter les relations fonctionnelles cachées entre les attributs, c'est-à-dire des relations imperceptibles pour l'estimateur (Bode, 2000).

En faisant référence à Lippmann (Lippman, 1987), Bode (2000) énonce que, comparés aux méthodes conventionnelles, les réseaux de neurones présentent beaucoup plus d'avantages. Ils font moins de sous-entendus des formes de distribution sous-jacentes que les méthodes statistiques traditionnelles et peuvent ainsi être plus robustes quand les distributions ne sont pas gaussiennes. Ils peuvent facilement rouler sur des processeurs en parallèle grâce au parallélisme de leur architecture, ce qui est négligeable pour la présente recherche, car la quantité de données n'est pas très grande. Ainsi, le temps de calcul est diminué. Ils fournissent un bon rendement de robustesse ou de tolérance à l'erreur, car il y a plusieurs nœuds ayant chacun une connexion locale. Quelques nœuds ou liens endommagés ne détériorent pas la performance de manière significative.

Il est important de souligner les observations de Bode concernant les données réelles ou les données artificielles dans le comportement des réseaux de neurones. Bode déclare que le point de vue d'un praticien serait que l'étude pilote basée sur des données réelles donnerait plus d'information quant à l'application des réseaux de neurones. Il est toutefois intéressant de compléter cette approche avec une simulation et en utilisant des données artificielles pour plusieurs raisons. Il est difficile de trouver une entreprise prête à ouvrir ses dossiers confidentiels contenant des données de coûts à un chercheur externe. La pertinence des réseaux de neurones dépend souvent de certaines caractéristiques d'application du domaine, comme le nombre de variables indépendantes ou le degré de similarité et le nombre de cas antérieurs. Il faudrait probablement beaucoup de temps avant d'identifier les bons domaines et d'obtenir les données réelles. Dans les situations réelles d'estimation de coût, la relation

fonctionnelle entre les propriétés du projet et le coût est inconnue. En conséquence, mis à part le coût antérieur, il n'y a aucun standard connu pouvant être utilisé pour mesurer la performance du réseau de neurones. Le nombre de données pour la formation et la mesure de la performance (test) est limité au nombre de cas antécédents

Les données artificielles sont plus pratiques pour explorer les architectures des différents réseaux, en étudiant le comportement du réseau sur des paramètres et des changements de données contrôlées. Cette observation devient intéressante si l'objectif d'une recherche future est de vérifier les performances des réseaux de neurones. Il faut néanmoins tenir compte de cette observation quand on vérifie l'application d'une prédiction par réseau de neurones dans un domaine réel.

Les résultats avec les données pilotes de cette recherche (Bode, 2000) suggèrent la compatibilité des réseaux de neurones pour le coût d'estimation de tâches de complexité similaire si le nombre de variables indépendantes n'est pas au-dessus de cinq ou six, et si au moins 50 à 100 projets sont disponibles comme historique de données.

D'un point de vue du développement d'un nouveau produit (manufacturier), les réseaux de neurones semblent plus appropriés pour le coût d'estimation dans la phase de design. Le nombre de variables indépendantes doit être de 5 ou 6, ce qui implique que les réseaux de neurones ne sont pas le choix approprié pour le coût d'estimation d'innovations radicales et de produits très complexes et pour les phases tardives de design lorsque les détails de la planification manufacturière sont bien connus. Pour ce qui est de la phase où, justement, les détails de la planification manufacturière sont connus, la régression (méthode paramétrique) n'est pas non plus appropriée. On devra se tourner vers des méthodes d'estimation déterministes.

Les trois recherches (Bode, 2000 Li *et al.*, 2001 Smith et Mason, 1996) concluent que les réseaux de neurones ont des avantages certains sur les régressions. Fait important à constater : Bode conclut que le nombre de variables indépendantes ne devrait pas dépasser cinq à six et que Li *et al.* en ont utilisé dans leur recherche un nombre inférieur à six. Il en est

de même pour la recherche de Smith *et al.* Devrait-on en conclure que cela est une caractéristique indispensable? On croit que cet état de fait est important et qu'il serait essentiel de l'approfondir. Bode dit qu'il est vital, pour les réseaux de neurones, d'avoir un échantillon d'au moins 50 à 100 données d'historique. Li *et al.* ont disposé d'un patron d'échantillon de plus de 1 500 unités, ce qui est en ligne avec les conclusions de Bode. Cependant, Smith *et al.* ont bénéficié de 16 échantillons. C'est peut-être ce qui explique des résultats moins probants.

Autre observation qui fait l'unanimité dans les trois recherches : les réseaux de neurones performant mieux que les méthodes de régression si les données d'entraînement sont non linéaires et qu'il y a des discontinuités fonctionnelles significatives.

En conclusion, si les différents facteurs énoncés précédemment tels que la non-linéarité des données entre les variables indépendantes ou les variables conductrices, s'il y a quelques données connues des caractéristiques de la valeur à prédire, si l'on est en partie certain que les attributs de l'historique de données sont partiellement en ligne avec les valeurs à prédire et si l'on n'est pas certain de ce qui influence la valeur à prédire, les réseaux de neurones sont plus performants que les méthodes de régression linéaire.

Comme les attributs d'entrée, des réseaux de neurones s'appuient sur des données que les estimateurs doivent de toute façon utiliser pour la préparation de leurs estimations, la technique des RNA semble appropriée comme algorithme de prédiction pour l'attribution du taux de productivité d'une équipe d'excavation de tranchée.

2.3 Résumé de l'évolution des réseaux de neurones dans le domaine de l'estimation des coûts de projet

Un résumé de l'évolution des réseaux de neurones dans le domaine de l'estimation des coûts est présenté ici :

- En 1960, les résultats donnent des perturbations dans les poids engendrant des imprécisions et une incapacité de généralisation des réseaux de neurones(Lu, Yeung et Ng, 2005).
- En 1962, Widrow et Hoff (IRE Wescon Convention Record) introduisent le concept de la sensibilité dans le réseau de neurones Adaline (réseau de neurones à une couche) et développent un réseau de neurones à une couche utilisant l'approche géométrique de n-dimensions(Lu, Yeung et Ng, 2005).
- En 1989, Winter (Stanford University) définit une expression analytique de la probabilité de l'erreur dans Madeline (réseau de neurones à couches multiples). Il explore le respect de la perturbation du poids(Lu, Yeung et Ng, 2005).
- En 1990, Stevenson (IEEE Transact Neural Net) poursuit l'étude de Winter et établit le concept de sensibilité dans l'erreur du poids(Lu, Yeung et Ng, 2005).
- En 1993, McKim (Cost Engineering) applique un réseau de neurones dans l'estimation des coûts et compare les résultats à trois méthodes d'estimation conventionnelles. Il arrive à la conclusion que les réseaux de neurones ont un bon potentiel pour l'estimation non déterministe(McKim, 1993).
- En 1994, Hegazy et Moselhi (*Journal of Computing in Civil Engineering*) utilisent un modèle évolué de réseau de neurones avec paradigme de rétropropagation. Ils comparent un modèle de réseau simple et un modèle hiérarchique. Le premier donne des résultats supérieurs, mais il y a encore beaucoup trop d'erreurs. Néanmoins, ils observent que la

capacité de généralisation des RNA est plus grande avec la technique de l'algorithme génétique, car celle-ci donne de meilleurs résultats. L'algorithme génétique permet de calculer le nombre optimal de couches cachées du réseau de neurones.(Adeli, 2001 Hegazy, Fazio et Moselhi, 1994)

- En 1995, Garza et Rouhana (Cost Engineering) étudient le potentiel d'application des réseaux de neurones en comparant les résultats à l'estimation paramétrique des coûts de travaux. Les réseaux de neurones donnent des résultats supérieurs et démontrent une solide capacité prédictive. Cependant, ils observent une complexité de la conception architecturale du réseau de neurones à cause des paramètres difficiles à calibrer : beaucoup d'essais / erreurs(Kim, An et Kang, 2004).
- En 1995, Creese et Li (Cost Engineering), à partir de données d'estimation paramétrique pour des ponts de bois, comparent les résultats d'un réseau de neurones avec la régression linéaire (Creese et Li, 1995).

Cet article veut démontrer qu'il est possible de transformer une série de données (coût total des travaux) en un réseau de neurones d'un degré de précision acceptable à partir duquel on peut établir des prédictions. Il cherche également à valider la précision de ce réseau de neurones en le comparant à la méthode standard de la régression linéaire.

Les données utilisées dans le réseau de neurones reposent sur la méthode paramétrique d'estimation, parmi les plus utilisées dans l'industrie. De plus, parvenir à l'objectif fixé dans cet article aurait pour effet d'utiliser, pour l'estimation, la méthode paramétrique au lieu de la méthode ascendante, ce qui représente une économie de temps considérable pour l'estimation des coûts, un des buts de la présente recherche.

Les résultats obtenus sont assez concluants. Le réseau de neurones est plus précis que la méthode de la régression linéaire. À l'intérieur du réseau de neurones, trois modèles ont été développés. Le troisième, soit le modèle utilisant toutes les données d'entrée, s'est avéré plus précis. Plusieurs tableaux résumant les différents résultats sont présentés

clairement dans l'article et il est assez simple de suivre l'évolution de la méthode. Par contre, étant donné que la méthode PlaNet est utilisée pour la construction du réseau de neurones et que cette méthode n'est pas décrite, il devient difficile de suivre ou de retrouver d'où viennent les valeurs des matrices obtenues.

Un des objectifs était de démontrer qu'il est possible de transformer une série de données (coût total des travaux) en un réseau de neurones d'un degré de précision acceptable, et à partir duquel on peut établir des prédictions. Cet objectif est atteint mathématiquement par les auteurs, et ce, en fonction d'un type de travail donné. Les résultats obtenus sont prometteurs et plusieurs autres études devront être effectuées afin de raffiner l'utilisation du neurone en fonction des différents problèmes rencontrés dans la construction. Cependant, il est important de se rappeler que la façon dont les paramètres de comparaison ont été choisis n'a fait l'objet de discussion.

L'article démontre, avec les trois modèles utilisés dans la méthode des réseaux de neurones, que le nombre de données indépendantes d'entrée a une influence importante sur la précision des résultats de prédiction. Plus le nombre de données est grand, plus la précision est bonne. Qu'en est-il du choix de ces données? Cet article n'en fait aucunement allusion.

L'article avait également comme objectif de valider la précision du réseau de neurones en le comparant à la méthode standard de la régression linéaire. Objectif totalement atteint, car il est clairement démontré que la précision du réseau de neurones est plus grande qu'avec la méthode de régression linéaire.

La problématique, énoncée en début d'article, est que la méthode ascendante en estimation des coûts de travaux en est une fastidieuse et dispendieuse et qu'il serait souhaitable de trouver une façon de substituer la méthode ascendante à la méthode paramétrique, plus rapide et moins coûteuse. Bien que cet article ne règle pas cette problématique, il permet d'amorcer une réflexion profonde sur l'utilisation de la méthode paramétrique dans l'estimation des coûts de travaux. Si ce réseau de neurones était validé

pour plusieurs autres types de donnée tels que temps de réalisation de travaux, la production d'équipe de travail, la formation d'équipe en fonction de type de travaux, etc., ce serait un grand pas pour une amélioration dans le processus d'estimation des coûts de construction.

- En 1997, Smith et Mason (*The Engineering Economist*) examinent la performance, la stabilité et la facilité de la modélisation de l'estimation des coûts en comparant la régression linéaire et un réseau de neurones. Ils découvrent que le réseau a un avantage si les données utilisées ont seulement un léger rapprochement avec les projets estimés. Par contre, la régression linéaire a un avantage considérable si les données utilisées sont totalement alignées avec le projet concerné (Smith et Mason, 1996).
- En 1998 et 2000, Bode (*Cost Engineering* 1998 et *International Journal of Production Research*) effectue une investigation expérimentale de l'application des réseaux de neurones lors de l'estimation des coûts en début du projet. Il compare le réseau de neurones à la régression linéaire et non linéaire. Il arrive à la conclusion que les réseaux de neurones donnent de meilleurs résultats (Bode, 1998, 2000).
- En 1998, Hegazy et Ayed (*Journal of Construction Engineering et Management*) tentent de surmonter les inconvénients des réseaux de neurones et présentent une approche simple et efficace pour les estimations paramétriques (Kim, An et Kang, 2004). Ils planifient un réseau de neurones à trois couches simulé dans un format de feuilles de calcul (Kim, An et Kang, 2004) facile à utiliser, transparent et familier pour beaucoup d'intervenants de la construction. Pour trouver une solution à la formation d'un réseau de neurones qui convient au modèle de la feuille de calcul, deux feuilles de calcul accompagnatrices ont été utilisées pour optimiser le poids du réseau de neurones basé sur la transmission unidirectionnelle et les procédures d'algorithme génétique. En utilisant le réseau de neurones à base de feuilles de calcul, un système complet pour l'estimation paramétrique des routes a été développé. Pour la configuration du réseau de neurones, Hegazy et al. proposent qu'une seule couche cachée est suffisante pour générer un lien arbitraire entre les entrées-sorties et que le nombre de neurones dans la couche cachée est

0,75 m, m ou 2 m +1, ou m est le nombre d'entrées dans le neurone. Pour cette étude, les trois valeurs proposées pour le nombre de neurones dans la couche cachée ont été utilisées (Hegazy, Fazio et Moselhi, 1994).

- En 1998, Yeh (*Journal Engineering and Management*) propose un réseau de neurones logarithmique et examine son efficacité et son exactitude pour l'évaluation des quantités d'acier dans le béton armé de bâtiment. Il démontre que ce type de réseau de neurones améliore significativement la performance dans la prise de quantités en acier d'armature (Kim, An et Kang, 2004).
- En 2004, Kim, An et Kang (*Science Direct Building and Environment*) comparent trois techniques, soit les réseaux de neurones, les *case-based reasoning* (CBR, voir Figure 2.1) et la régression linéaire. Ils arrivent à la conclusion que le réseau de neurones est plus précis, mais que le CBR donne de très bons résultats et qu'il est beaucoup plus simple à maîtriser. Ils recommandent pour les futures recherches des modèles hybrides (Kim, An et Kang, 2004).

G.-H. Kim et al. / *Building and Environment* 39 (2004) 1235–1242

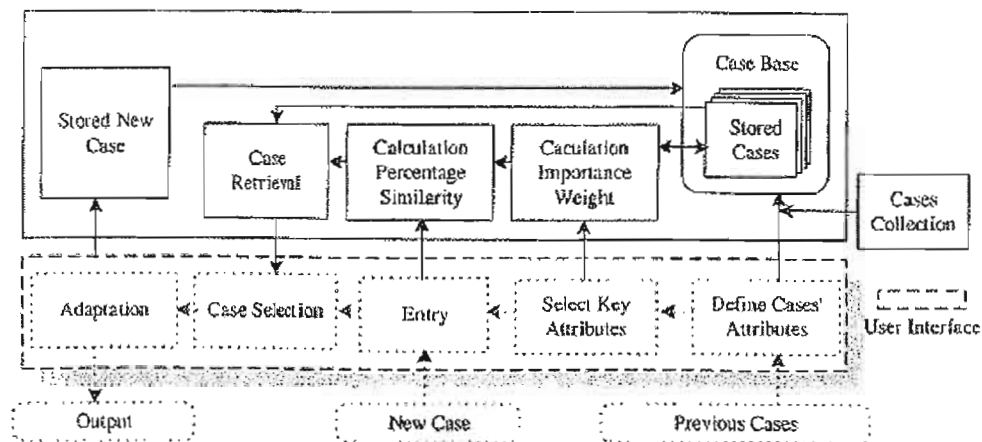


Fig. 2. Case-based reasoning model.

Figure 2.1 : *Case-based reasoning* (CBR) (Tirée de G.H. Kim et Al, 2004.)

- En 2004, Kim, Yoon, An, Cho et Kang (Science Direct Building and Environment) découvrent que les réseaux de neurones éliminent l'obligation de trouver un historique de données totalement aligné mathématiquement avec le projet à estimer et que ces réseaux de neurones n'ont pas de restrictions sur le nombre de variables de coûts à cause de la possibilité qu'ils ont de s'organiser et se bonifier d'eux-mêmes. Ils découvrent que la structure du vecteur de tendance n'a pas à être déterminée préalablement, bien que le réseau de neurones d'activation doive être choisi au départ. Ils observent également que si le nombre de couches cachées est insuffisant, le réseau de neurones ne fait pas de liens entre les intrants et les extrants, et si le nombre de couches est trop élevé, malgré qu'on augmente la précision, les performances informatiques sont exécrables. Ils concluent également qu'étant donné l'inexistence de règle appropriée pour obtenir la taille optimale du réseau et que la configuration de l'ensemble des paramètres de rétropropagation est nécessaire lors de la phase d'apprentissage du réseau de neurones ceci engendre beaucoup d'essais-erreurs (Kim *et al.*, 2004).
- En 2004, Kim, Yoon, An, Cho et Kang (Science Direct Building and Environment) explorent deux méthodes hybrides. Un réseau de neurones avec rétropropagation, où le nombre de couches est calculé par tâtonnement, est comparé à un réseau de neurones avec rétropropagation auquel ils ont introduit un algorithme génétique (voir Figure 2.2) pour la détermination du nombre de couches. Le réseau avec l'algorithme génétique est plus performant et diminue de beaucoup le temps et les efforts pour optimiser le nombre de couches du réseau de neurones.

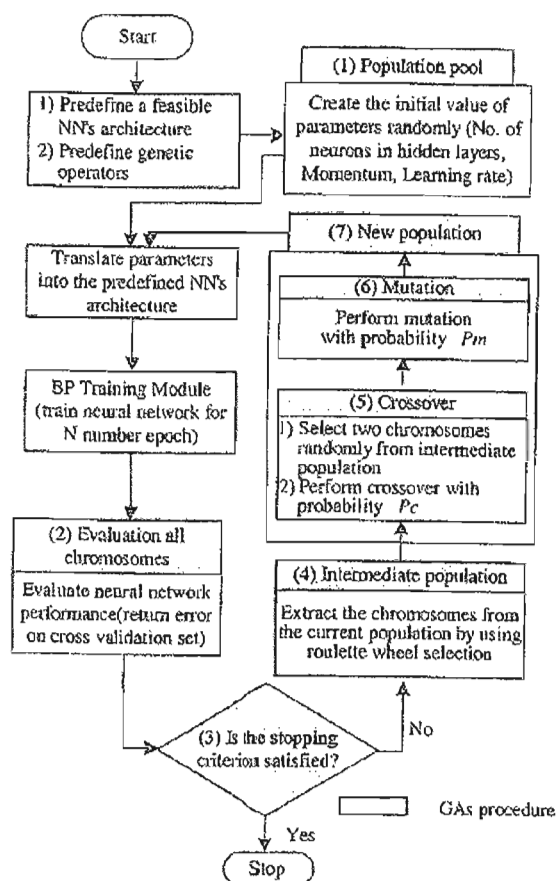


Fig. 1. Using GAs to optimize parameters of BPN.

Figure 2.2 : Algorithme génétique (Tirée de G.H. Kim et al., 2004.)

Le 20 avril 2005, Lu, Yeung et Ng (Soft computing) analysent la sensibilité d'un perceptron à plusieurs couches (MLP-NN) en fonction des intrants avec des unités de mesure différentes. La vérification du modèle MLP-NN est effectuée sur une production d'une équipe de coulée de béton pour la construction d'un bâtiment. Ils obtiennent de très bons résultats. La production fut de 21,65 mètres cubes / heure en estimation et de 21,35 mètres cubes / heure au chantier. Ils suggèrent de tester le modèle sur de nouveaux scénarios dans le domaine de l'ingénierie (Lu, Yeung et Ng, 2005).

2.4 Conclusion

En résumé, les chercheurs concentrent leurs recherches sur les réseaux de neurones surtout autour des qualités telles que la précision, la sensibilité, la performance (coût en temps), la facilité d'utilisation et la capacité de généralisation. Ils les vérifient sur des estimations paramétriques à partir de projets, de produits, etc., où l'information n'est pas disponible en totalité (période de conception non terminée).

McKim R. (1993), Garza J. *et al.* (1995), Creese R. *et al.* (1995), Li H. (1995), Smith AE *et al.* (1997), Adeli H. *et al.* (1998), Bode J. (1998 et 2000) et Yeh IC (1998) ont exploré l'application des réseaux de neurones afin d'améliorer la précision des estimations paramétriques des coûts au-delà des limites du modèle de la régression linéaire. Les recherches précédentes ont révélé que les réseaux de neurones étaient supérieurs au modèle de la régression linéaire dans l'estimation des coûts.

En 1994, Hegazy et Moselhi (*Journal of Computing in Civil Engineering*) observent que la capacité de généralisation est plus grande avec la technique de l'algorithme génétique en donnant de meilleurs résultats. En 2004, Kim, Yoon, An, Cho et Kang (Science Direct Building and Environment) concluent que le réseau avec l'algorithme génétique est plus performant et diminue de beaucoup le temps et les efforts afin d'optimiser le nombre de couches du réseau de neurones.

On peut conclure que les chercheurs ont beaucoup travaillé sur la paramétrisation des réseaux de neurones afin de diminuer le temps et faciliter la tâche pour l'apprentissage de ces réseaux. Les travaux de recherche sur les infrastructures civiles ont été concentrés en général sur l'estimation paramétrique utilisée lorsque le degré de définition du projet se situe dans les classes 4 et 3. Ce qui signifie que le projet n'est pas trop avancé. Les conclusions de ces recherches démontrent que les résultats obtenus sont assez probants et méritent d'être utilisés.

Par contre, on a démontré que pour l'estimation détaillée, l'attribution du taux de productivité d'une équipe de travail est une source d'erreur importante. Certains chercheurs (Bode, 2000)

croient que les réseaux de neurones ne sont pas destinés pour les projets où le degré de définition est très avancé et que l'estimation détaillée est plus appropriée. Malgré cette affirmation, on croit que les réseaux de neurones pourraient être utiles pour certaines tâches de la méthode d'estimation détaillée telle que l'attribution du taux de productivité d'une équipe de travail. Cette recherche vise à explorer cette hypothèse.

L'attribution du taux de productivité d'une équipe de travail a certaines similitudes avec les plateformes sur lesquelles les essais sur les estimations paramétriques ont été effectués.

Il est certain que les chercheurs sont en majorité arrivés à la conclusion que les RNA sont plus utiles lorsque le niveau de définition du projet est peu élevé. C'est une conclusion propre au domaine, car la majorité des recherches tend à prédire le coût des travaux avec une meilleure précision que la précision des estimations paramétriques dont la justesse du coût final du projet selon l'AACE est de l'ordre de 60 % / -30 %. Améliorer de façon significative cette précision sans avoir à augmenter le niveau de définition du projet serait un plus significatif pour l'ensemble de l'industrie de la construction. Par contre, il y a un fait que ne faut pas nier : les estimations détaillées seront toujours des méthodes d'estimations utilisées pour la préparation des soumissions que les entrepreneurs doivent faire pour l'obtention d'un contrat de construction d'un projet. Elles doivent être le plus précises possible, car des imprécisions peuvent engendrer des pertes monétaires considérables. Trouver des méthodes pour améliorer cette précision serait également un plus significatif pour l'industrie de la construction.

Les similitudes de l'attribution du taux de productivité avec les plateformes sur lesquelles les essais antérieurs ont été effectués se trouvent autour du fait et de la démonstration effectuée au Chapitre I. Il a été démontré que le taux de productivité était tributaire de plusieurs facteurs entremêlés qui évoluaient de façon indépendante ou de façon liée entre eux. Cela veut donc dire que le niveau de définition du taux de production est peu élevé et il est même difficile de mesurer son évolution. Selon la majorité des conclusions des recherches, cela est le genre de cas où les réseaux de neurones performant bien.

À première vue, le nombre de données d'entrée qui influencent le taux de productivité est de l'ordre d'une quinzaine. Malgré que le montant soit un peu élevé, ce nombre rejoint les recherches subséquentes sur les estimations paramétriques. Après quelques essais, il serait peut-être possible de diminuer ce nombre en se basant sur les conclusions de la recherche de Setyawati et Creese (Setyawati, Creese et Sahirman, 2003; Setyawati, Sahirman et Creese, 2002). Selon leurs résultats, éliminer certaines données d'entrée n'a pas d'impact significatif sur la précision de prédiction.

CHAPITRE III

LES DONNÉES BRUTES

3.1 Généralité

La présente recherche concerne l'exploration de la capacité des réseaux de neurones à imiter le jugement et l'expérience d'un estimateur chevronné. Pour atteindre cet objectif et pour le paramétrage des RNA, on se base sur les conclusions des ouvrages de Kim *et al.*, de Lu *et al.* et de Hegazy *et al.* sur les réseaux de neurones avec rétropropagation.

Étant donné que cette recherche s'appuie sur des théories existantes et acceptées, dont l'objectif est d'expliquer et de vérifier les résultats obtenus, l'axe méthodologique se veut hypothético-déductif. Il s'agit d'une recherche à caractère exploratoire dont la méthode est quantitative.

Selon Cooper et Schindler (Cooper, 2002), au Chapitre 2, la méthode scientifique de recherche se base sur les principes fondamentaux de la science tels que l'observation directe d'un phénomène, les variables, les méthodes et les procédures clairement définies, une hypothèse empirique testable, la capacité d'exclure les hypothèses opposées. Toujours selon Cooper et Schindler (2002), certains chercheurs utilisent souvent une question de recherche au lieu d'une hypothèse descriptive.

Les données de cette recherche proviennent de projets de génie urbain exécutés par plusieurs entreprises de construction d'infrastructures civiles de la région de la rive nord de Montréal sous la surveillance d'une firme de génie-conseil. Ces dernières sont toutes des entreprises d'expérience, en affaires depuis plus de 20 ans. Certaines utilisent également des logiciels commerciaux pour l'estimation, la maîtrise des coûts de projet et la comptabilité. Elles réalisent en général des travaux de génie urbain pour des projets situés en majorité sur la rive nord de Montréal. Pour la surveillance des travaux, les firmes de génie affectent une ressource appelée surveillant de chantier. Dans le cadre de ces fonctions, un surveillant doit compléter un rapport journalier qui résume l'ensemble des travaux exécutés dans la journée. C'est à partir de ces comptes rendus que les taux de productivité journaliers ont été répertoriés dans l'historique de données après un certain traitement des données brutes.

La méthodologie suivie est celle suggérée dans l'ouvrage de Han et Kamber (Han, 2001) et les différentes étapes sont la définition du problème, l'identification et la préparation des données, le choix d'un échantillon, la construction et le testage du modèle, l'évaluation du modèle et le choix des paramètres optimaux, l'application du modèle à l'ensemble des données et l'interprétation des résultats.

Dans ce chapitre, on développe les différentes étapes à partir de la définition du problème jusqu'au choix d'un échantillon. Les autres étapes seront décrites dans le chapitre suivant. Le but de séparer l'ensemble de ces étapes en deux chapitres différents est de faire ressortir les éléments appartenant au domaine de la construction de ceux du domaine de l'informatique. En d'autres mots, ce chapitre décrit les étapes dont l'information émane des experts du domaine de la construction, tandis que le suivant couvre la partie de traitement des données, d'entraînement du RNA et de l'interprétation des résultats qui représente la partie informatique de la recherche.

3.2 La définition du problème

La définition du problème concerne les objectifs de cette recherche : explorer la capacité des réseaux de neurones à imiter le jugement et l'expérience d'un estimateur chevronné. On compare les taux de productivité d'une équipe de travail obtenus par un RNA avec les taux de productivité réels. Puis, on compare les taux de productivité déterminés par trois estimateurs avec le taux de productivité réel. Les pourcentages d'erreurs des résultats obtenus avec les RNA sont comparés avec les pourcentages d'erreurs des estimateurs chevronnés.

Étant donné que la régression linéaire a donné d'excellents résultats dans certains cas, on compare les performances des RNA avec les performances de la régression linéaire.

Un des buts de la recherche est de noter les performances informatiques en coût temps des RNA. Ces lectures pourront servir de base pour des recherches futures afin d'évaluer la possibilité d'intégrer des algorithmes de prédiction à des logiciels commerciaux d'estimations détaillées des coûts de travaux.

3.3 Identification et préparation des données brutes

Toutes les données sont recueillies à partir des rapports journaliers des surveillants des travaux d'une firme de génie-conseil membre de l'AICQ. Ces travaux en sont d'infrastructures municipales totalisant des coûts de projet variant de 250 000 @ 5 000 000 \$ exécutés sur le territoire de Ville de Laval ou sur celui des Basses-Laurentides. Soixante-quinze projets ont été sélectionnés pour une analyse préliminaire.

Par manque d'information ou à cause des contradictions dans les informations recueillies aux rapports journaliers, seulement 43 projets ont été retenus, à l'intérieur desquels 412 rapports journaliers ont été jugés pertinents pour cette recherche représentant 60 % des rapports totaux.

Plus précisément, les données répertoriées sont le taux de productivité des équipes de travail et les ressources en matériel, en main-d'œuvre et en équipement, ainsi que la nature des sols rencontrés pour la pose de conduite d'aqueduc, de conduite d'égout sanitaire et de conduite d'égout pluvial incluant les différents accessoires tels que regards d'égout, vanne et chambre de vanne, borne-fontaine, puisard et entrée de service. C'est travaux sont appelés communément dans le domaine de la construction « travaux de tranchée ».

Les données totalisent 141 tranchées d'une conduite (en général l'aqueduc), 152 tranchées de deux conduites (en général l'aqueduc et l'égout sanitaire), puis finalement 118 tranchées de trois conduites.

Par la suite, le logiciel commercial GESQAN a été utilisé pour calculer le volume théorique d'excavation ainsi que le volume de remblai à partir des plans et devis de chacun des projets. Pour les projets exécutés en terrain vague, les tranchées ont été considérées remblayées avec les matériaux d'excavation jusqu'au niveau du terrain naturel, tandis que celles effectuées sur des rues existantes ont été considérées remblayées avec les matériaux d'excavation jusqu'au niveau de l'infrastructure du pavage existant. Le Tableau 3.1 ci-après présente les données sources des rapports journaliers d'inspection de six tranchées d'un même projet. Il existe un tableau semblable pour l'ensemble des 43 projets totalisant 412 journées de travail d'excavation de tranchée.

Tableau 3.1
Données sources des rapports journaliers d'inspection

Tranchée	TR842001	TR842002	TR842003	TR842004	TR842005	TR842006
Date d'exécution	10 / 12 / 2005	11 / 12 / 2005	13 / 12 / 2005	14 / 12 / 2005	30 / 01 / 2006	31 / 01 / 2006
Heure de travail	07 h-18 h 30	07 h-17 h	07 h-17 h	07 h-15 h 30	07 h-16 h 30	07 h-15 h 30
Durée travail (h)	11,0	9,5	9,5	8,0	8,5	7,5
Temp. (°C)	-4	-2	-18	-18	-5 / 2	-4 / 1
Météo	Nuageux	Nuageux	Ensoleillé	Ensoleillé	Pluvieux et nuageux	Pluvieux et nuageux
Chaînage	1+002 à 1+011	1+011 à 1+016	1+020 à 1+024	1+024 à 1+030	1+166.8 à 1+174.8	1+174.8 à 1+178.8
Longueur par jour (m)	9,00	5,00	4,00	6,00	8,00	4,00
Longueur par heure (m / h)	0,82	0,53	0,42	0,75	0,94	0,53
Productivité (GPL) (ml / h)	1,00	0,59	0,42	0,86	1,07	0,53
Volume d'excav. (m3)	98,35	54,64	44,35	63,35	70,86	34,06
Productivité Pelle (m3 / h)	8,94	5,74	4,67	7,92	8,34	4,54
Élévation moy. du T.N.	29,350	29,350	29,400	29,400	28,900	28,900
Nature du sol	Roc, terre	Roc, terre	Roc, terre	Roc, terre	Roc, terre	Roc, terre
Pente d'excavation	Moyenne	Moyenne	Moyenne	Moyenne	Moyenne	Moyenne
Élévation du roc	-----	-----	-----	-----	-----	-----
AQ (PVC)	200	200	200	200	200	200
E.S. (PVC)	250	250	250	250	250	250
E.P. (TBA)	-----	-----	-----	-----	300	200
Écart (m) AQ / E.S.	1.5	1.5	1.5	1.5	1.5	1.5
Écart (m) E.S. / E.P.	-----	-----	-----	-----	1.5	1.5
Élev. moy. du pavage prop.	29,350	29,350	29,400	29,500	29,400	29,400
Couvert moy. AQ. (m)	2,00	2,00	2,00	2,00	2,00	2,00
Couvert moy. E.S. (m)	2,95	2,95	3,00	3,00	2,50	2,50
Couvert E.P. (m)	-----	-----	-----	-----	1,50	1,50
Nombre Regard	0	1	0	0	1	0
Nombre Vanne	1	0	0	0	0	0
Nombre BF	0	0	0	0	0	0
Nombre CV	0	0	0	0	0	0
Nombre Puisard	0	0	0	0	0	0

Tableau 3.1 (suite)
Données sources des rapports journaliers d'inspection

Tranchée	TR842001	TR842002	TR842003	TR842004	TR842005	TR842006
Raccord, à l'existant	1	0	0	0	0	0
Entrée de services	0	0	0	1	0	0
Objets installés	V6	RS1	-----	-----	RP4	-----
Nombre de journaliers	2	2	2	2	2	2
Type de machinerie	2 pelles mécaniques (Tramars 892ELC [non utilisé] et John Deere 330LC) 1 pépine (N / A) 1 jumping jack	2 pelles mécaniques (Tramars 892ELC [n-u] et John Deere 330LC) 1 pépine (N / A) 1 chargeur (N / A) 1 camion (10 roues) 1 jumping jack	2 pelles mécaniques (Tramars 892ELC [n-u] et John Deere 330LC) 1 pépine (N / A)	2 pelles mécaniques (Tramars 892ELC [n-u] et John Deere 330LC) 1 pépine (N / A) 1 chargeur (N / A) 1 camion (10 roues) 1 jumping jack 1 plaque vibrante	2 pelles mécaniques (Tramars 892ELC [n-u] et John Deere 330LC) 1 pépine (N / A) 1 bélier (N / A) 1 chargeur (N / A) 2 camions (semi) 1 jumping jack 1 plaque vibrante	2 pelles mécaniques (Tramars 892ELC [n-u] et John Deere 330LC) 1 pépine (N / A) 1 bélier (N / A) 1 chargeur (N / A) 2 camions (semi) 1 jumping jack 1 plaque vibrante
Type de chantier	1	1	0	0	0	0

Un code est attribué à toutes les tranchées pour fin d'identification. Il est composé des lettres TR pour tranchée suivie d'un chiffre de trois nombres représentant le numéro du projet d'où émane la tranchée ainsi que de trois chiffres représentant celui de la conduite. À titre d'exemple, la première tranchée du projet 842 est identifiée avec le code suivant : TR842001. Les numéros de projet sont ceux utilisés par la firme de génie.

3.4 Choix de l'échantillon selon les experts du domaine

Il est certain que pour déterminer les facteurs influençant le taux de productivité, on pourrait se fier à la recherche de Rowings (1996) qui définit trois catégories de facteurs tels que ceux reliés à la gestion, ceux reliés au projet et ceux reliés aux tâches. Cependant, il devient intéressant de consulter des praticiens chevronnés actifs dans l'industrie de la construction du génie urbain et faisant quotidiennement des soumissions à partir de la méthodologie d'estimations détaillées du coût de projet dans le cadre des appels d'offres des donneurs d'ouvrage. Une consultation a été organisée auprès de trois estimateurs chevronnés de l'industrie ayant plus de 25 années d'expérience dans ce domaine. Afin de ne pas influencer leur jugement, chacun a été questionné individuellement. Lors de cette rencontre, il a été discuté des objectifs de cette recherche et on leur a demandé de déterminer selon eux quels sont les éléments qui influencent le taux de productivité d'une équipe qui procède à la construction de réseaux d'aqueduc et d'égout.

Les facteurs d'influence du taux de productivité selon l'estimateur n° 1 sont les suivants :

1. Capacité de production de la pelle / pelletage en fonction du volume totale qui a une certaine influence.
2. Nombre de conduite à installer dans la tranchée (grande importance).
3. Type et diamètre des tuyaux (PVC, Hyprescon, béton, fonte, etc.).
4. Profondeur maximum en fonction de la portée maximale de la pelle (elle varie de cinq @ huit mètres).
5. Méthode de *casting* (au bout de la portée de la pelle, de côté ou à l'arrière).

6. Travaux à exécuter dans un terrain vacant ou dans des infrastructures municipales existantes.
7. Travaux d'entrée de service fait en même temps que les travaux de la tranchée principale ou de façon séparée.
8. Présence d'utilités publiques.
9. Importante de la saison seulement si cela améliore les conditions.
10. Carnet d'ouvrage de l'entrepreneur : s'il n'est pas rempli, la productivité est plus faible.
11. Longueur totale du projet à effectuer : plus il dure, plus la productivité est meilleur et vice-versa).
12. Entrée de services à passer sous la conduite principale.
13. Type de sol à excaver.
14. Optimisation de l'équipe de travail (méthode de travail).

Les facteurs d'influence du taux de productivité selon l'estimateur n° 2 sont les suivants :

1. Le nombre de conduite à installer dans la tranchée.
2. Diamètre des tuyaux à installer dans la tranchée.
3. Type de conduite à installer dans la tranchée (PVC, fonte, béton, Hyprescon, etc.).
4. Profondeur à laquelle les conduites seront installées.
5. Pente à laquelle les conduites seront installées.
6. Type de sol dans lequel on excave (pente d'excavation),
7. Niveau de la nappe phréatique (pompage).
8. Type de matériel avec lequel on doit remblayer la tranchée (camionnage).
9. Environnement physique (proximité des maisons, des routes, des fils aérien...).
10. Obstacles rencontrés dans la tranchée (Bell, gaz, massif, Hydro-Québec...)
11. Nombre d'accessoires à installer (entrées de services, puisards, regards, chambre de vanne...).
12. Type et nombre d'équipements lourds affectés aux travaux.
13. Expérience des contremaîtres, des opérateurs et journaliers affectés aux travaux.
14. Nombre de journaliers affectés aux travaux.

Les facteurs d'influence du taux de productivité selon l'estimateur n° 3 sont représentés au tableau ci-dessous.

Tableau 3.2
Les facteurs d'influence du taux de productivité selon l'estimateur no 3

Spécifications du projet	Caractéristiques du projet	Facteur d'influence	Commentaires
Site des lieux	<ul style="list-style-type: none"> Espace de travail 	<ul style="list-style-type: none"> Grosueur d'équipement possible Espace pour les opérations de l'équipement (fil aérien, conduits souterrains...) Entreposage du matériau d'excavation à côté des camions 	
Analyse des tests de sol	<ul style="list-style-type: none"> Roc, <i>boulder till</i> Argile Sable et eau, argile molle 	<ul style="list-style-type: none"> Difficile à excaver Facile d'excavation Pente instable, boîte de tranchée 	
Analyse des plans	<ul style="list-style-type: none"> Coupe type des sections à excaver 	<ul style="list-style-type: none"> Petit volume (inférieur à 10 m.cu. / m.lin.) Gros volume (supérieur à 10 m.cu. / m.lin.) 	<ul style="list-style-type: none"> Production limitée par la pose des conduites On peut augmenter le volume par jour en rajoutant un équipement
	<ul style="list-style-type: none"> Diamètre des conduites 	<ul style="list-style-type: none"> Inférieur à 120 mm de diamètre 1 200 mm de diamètre et plus 	<ul style="list-style-type: none"> Production normale Influence la production
	<ul style="list-style-type: none"> Nombre de conduites 	<ul style="list-style-type: none"> une conduite deux conduites trois conduites 	<ul style="list-style-type: none"> Influence le temps de pose et par le fait même la productivité
	<ul style="list-style-type: none"> Présence d'utilités publiques 	<ul style="list-style-type: none"> Si on croise ou longe un massif ou une conduite de gaz, cela aura automatiquement une influence sur la productivité 	
	<ul style="list-style-type: none"> Profondeur de la tranchée 	<ul style="list-style-type: none"> Inférieure à 6,0 mètres Supérieure à 6,0 mètres 	<ul style="list-style-type: none"> Aucun problème Prévoir palier ou grosse pelle (650)
	<ul style="list-style-type: none"> Type de remblai de la tranchée 	<ul style="list-style-type: none"> Dans un terrain vacant avec le même matériel ou remblai de classe A, mais on peut laisser le matériel sur le bord de l'excavation Remblai avec classe A ou matériau d'excavation, mais nous devons prévoir des camions 	<ul style="list-style-type: none"> Production normale Production avec camions

Tableau 3.2 (suite)
Les facteurs d'influence du taux de productivité selon l'estimateur no 3

Spécifications du projet	Caractéristiques du projet	Facteur d'influence	Commentaires
Analyse des plans	<ul style="list-style-type: none"> Type de conduite 	<ul style="list-style-type: none"> Si seulement une conduite d'aqueduc Si conduites d'égout sanitaire en PVC 	<ul style="list-style-type: none"> Moins de précision et de perte de temps Attention : précision et compaction importante
Divers	<ul style="list-style-type: none"> Équipement disponible (exemple : pelle 200 ou 650) Qualité de la main-d'œuvre disponible 	<ul style="list-style-type: none"> Influence sur la production 	

Les différents facteurs d'influence observés par les estimateurs consultés pourraient être regroupés en grande partie dans les différentes catégories déterminées par Rowings (1996). Cependant, on peut remarquer que les facteurs reliés aux tâches ne semblent pas préoccuper ces estimateurs. Après les avoir questionnés sur le sujet, ils ont été unanime à répondre que les facteurs tels que la motivation, le moral, la fatigue, le travail syndiqué, la qualité du travail bien fait, l'absentéisme et le taux de roulement étaient tous des facteurs qui ont un impact sur le taux de productivité, mais que la variation de ce taux est reflétée dans les projets antérieurs sur lesquels ils se basent pour définir le taux de productivité. C'est une bonne nouvelle, car moins on a attribut d'entrée pour les NN, plus est facile la configuration.

3.5 Conclusion

Nous avons vu dans ce chapitre d'où venaient les données brutes. Elles émanent des rapports journaliers des surveillants de travaux de génie urbain d'une firme de génie-conseil de la région de Laval détenant un certificat ISO-9001 : 2000. Bien que ces rapports journaliers soient vérifiés quotidiennement, il n'en demeure pas moins qu'il peut y avoir des erreurs, car la qualité de ces rapports dépend des connaissances et de la rigueur des surveillants et de leur supérieur immédiat qui les préparent. Ils représentent des projets ayant été exécutés sur plusieurs années, soit environ cinq ans. Le surveillant utilise une forme prédéfinie de ce rapport pour compléter le rapport journalier et, après cinq ans, ce rapport prédéfini n'a pas été modifié. Cela représente un avantage certains en ce qui concerne la consistance du contenu et la qualité de ces rapports.

Une observation importante est que les rapports sont montés de façon à quantifier l'ouvrage à exécuter pour une journée à la fois. Pour cette recherche, la donnée majeure répertoriée est le taux de productivité pour les travaux de tranchée. Une source d'erreur possible : lorsque l'équipe de travail a effectué des travaux différents durant la journée et qu'ils ont effectué également des travaux de tranchée. Pour obtenir le taux de productivité, on doit soustraire la durée des autres travaux effectués durant la journée. Par exemple, on retranche une heure et demie pour les travaux d'installation d'une borne-fontaine. Cela demeure une excellente approximation, car l'ensemble des experts consultés est unanime sur cette durée. Par contre,

la longueur de tranchée effectuée durant la journée est beaucoup plus courte et le taux de productivité, dans certain cas, donne des valeurs isolées des autres valeurs.

En ce qui a trait au choix de l'échantillon selon les experts, on obtient en général plus ou moins une quinzaine de facteurs influençant le taux de productivité d'une équipe de travaux de tranchée. Ce nombre pourra être révisé à la baisse car, après analyse, certains des facteurs énoncés sont tributaires d'autres facteurs dont on tient déjà compte.

CHAPITRE IV

TEST SUR LES RNA, RÉSULTATS ET INTERPRÉTATION

4.1 Généralité

Le chapitre précédent, a traité du travail effectué sur les données brutes du projet. On a également, avec les experts du domaine, analysé les facteurs influençant le taux de productivité d'une équipe d'excavation. Toutes ces analyses sont issues du domaine de la construction. Dans ce chapitre, on traite de la partie informatique. Dans l'ordre : la construction et le test du modèle, l'évaluation du modèle et le choix des paramètres, l'application du modèle à l'ensemble des données, et finalement, les résultats et leur interprétation. Le chapitre se termine par une conclusion sur les différents points observés dans les résultats afin de préparer la conclusion de cette recherche.

4.2 Construction et testage du modèle

Avant de choisir le moteur d'inférence, quelques logiciels ont été analysés : Matlab, SNNS et Weka. Le moteur d'inférence Weka a été utilisé parce qu'il est spécialisé en apprentissage machine, détient une multitude d'algorithmes disponibles et offre un API de développement complet en Java ou C# (.net). La plupart des logiciels commerciaux d'estimation sont développés justement en .net. De plus, il est gratuit sous la licence GNU et documenté avec un wiki.

Weka est une collection d'algorithmes d'apprentissage machine pouvant effectuer des tâches d'exploration de données. Les algorithmes peuvent être appliqués directement sur les données ou appelés d'une application programmée. Le logiciel inclut des outils de manipulation de données pour la préorganisation, la classification, la régression, le groupement en grappes, les règles d'association et la visualisation.

C'est en 1993 que l'Université de Waikato, en Nouvelle-Zélande, a commencé le développement de la version originale de Weka, devenu un mélange de TCL / TK, C et Makefile, des langages et scripts informatiques. En 1997, la décision de redévelopper Weka à partir de zéro en Java, incluant l'implémentation des algorithmes, a été prise. Quelques années plus tard, en 2005, Weka reçoit le Data Mining and Knowledge Discovery Service Award de l'Association for Computing Machinery (ACM).

Les buts, guides et règles de développement fixés par l'équipe de développement sont de rendre les techniques d'apprentissage machine le plus accessible possible, de les appliquer dans les problèmes importants de l'industrie de la Nouvelle-Zélande, de développer de nouveaux algorithmes d'apprentissage machine et de les mettre à la disposition du monde scientifique et de contribuer à un logiciel intégré (*framework*) pour le domaine.

Weka est développé et maintenu par l'équipe Weka de l'Université de Waikato et est encore très actif; la dernière mise à jour majeure a été publiée le 1^{er} juin 2007, soit la version 3.4.11. Par contre, un utilisateur peut télécharger la version de développement en cours (non officielle) du site de contrôle de version s'il veut expérimenter avec les tout derniers développements (cette version est 3.5.6).

Weka nécessite Java pour s'exécuter en mode autonome, mais si l'utilisateur veut incorporer ses fonctionnalités dans un logiciel indépendant, il a la possibilité d'utiliser les interfaces de programmation en Java ou C# (.net), deux langages très populaires. Ces interfaces sont très complètes, de haut et bas niveau (elles permettent donc une intégration totale) et documentées selon les règles JavaDoc de Sun Microsystems.

La machine de développement est un MacBook 2 GHz Intel Core 2 Duo, 1 Gb RAM. La machine de test ou à partir de laquelle l'apprentissage du réseau de neurones a été fait est un ordinateur AMD Athlon 64 3000+ de 1.81 GHz avec 1.25 Gb RAM. Le langage d'exploitation est celui de Microsoft Windows XP SP2. Et tel que décrit précédemment, le moteur d'inférence utilisée est le Weka 3.4.11 en mode autonome.

Avant d'utiliser les données sources, un traitement initial a été effectué afin de retirer quelques données, car elles n'étaient pas complètes ou elles ne permettaient pas un apprentissage adéquat. À titre d'exemple, si un seul projet a été effectué avec une conduite d'aqueduc 600 mm de diamètre par exemple, il aurait été alors très difficile de lui faire apprendre et estimer un futur projet. Cela sera détaillé ultérieurement.

Pour chaque essai d'apprentissage, l'ensemble de données, l'algorithme et les résultats sont présents sous son propre répertoire numéroté et contient des fichiers de types .CSV. Les classeurs Excel sont reconfigurés pour contenir les données de l'apprentissage sous un format facilement transférable en ARFF qui a aucun lien direct avec les fichiers d'apprentissage .RTF. Les documents textes contenant les résultats de l'apprentissage sans reformatage, soit directement du moteur d'inférence, sont sous la forme .ARFF. Pour références, ces fichiers contiennent chaque estimation pour chaque donnée. Les données formatées pour l'apprentissage et l'estimation sont conservées en fichiers qui contiennent un en-tête décrivant les données et leur format avec des clés, comme représenté au Tableau 4.1.

Tableau 4.1
Nom des fichiers des données formatées pour l'apprentissage du RNA

Code	Explication
% TRAINING DATA	Le % désigne un commentaire, ce qui suit le % n'est pas pris en charge par le moteur d'inférence.
@relation 'trenches'	La clé @relation désigne le nom de l'ensemble de données.
@attribute AQ real	La clé @attribute désigne le nom de l'attribut et le type. Le nom doit être une chaîne sans espaces et le type peut-être <i>numeric</i> , <i>nominal</i> , <i>string</i> ou <i>date</i> .
@data	Souligne que les lignes suivantes seront les données. Une série par ligne et chaque attribut sont séparés par une virgule. Le point désigne les décimales.

4.3 Évaluation du modèle et choix des paramètres optimaux

Les données d'entrée ont été choisies selon une priorité accordée en fonction de leur impact sur le résultat final du taux de productivité. La priorité est représentée au Tableau 4.2. Plus la priorité est petite, plus le paramètre a un impact sur la productivité de la tranchée. Ces priorités proviennent des discussions avec les experts du domaine consultés.

Tableau 4.2
Priorité des données d'entrée pour le taux de productivité d'une équipe de tranchée

Élément	Priorité
Aqueduc (PVC)	1
Égout sanitaire (PVC)	1
Égout pluvial (TBA)	1
Couvert moyen aqueduc (mètre)	2
Couvert moyen égout sanitaire (mètre)	2
Couvert égout pluvial (mètre)	2
Productivité pelle (mètre cube / heure)	3
Nature du sol	4
Regard	5
Couvert	5
Borne	5
Puisard	5
Vanne	5
Raccord	5
Entrée de service	5

Tableau 4.2 (suite)
 Priorité des données d'entrée pour le taux de productivité d'une équipe de tranchée

Élément	Priorité
Nombre de journaliers	5
Élévation du roc	6
Météo	7
Autres paramètres	8
Longueur par jour (m)	Résultat

Weka offre plusieurs méthodes pour tester l'apprentissage du RNA, soit de reprendre les mêmes données et de faire l'estimation, de prendre des données dans deux fichiers différents (séparées manuellement) ou de faire un choix au hasard par pourcentage (ex : 70 % / 30 %).

Pour ce projet, la deuxième option a été choisie et les données ont été séparées manuellement selon un choix logique décrit ultérieurement. Des difficultés ont été éprouvées puisque le nombre de données disponibles pour l'apprentissage est limité comme tenu du nombre de variables d'entrée. Le problème est qu'utiliser peu d'attributs dans l'historique de données ne permet pas à l'apprentissage d'établir des liens solides pour l'estimation. Des données identiques peuvent donner un résultat très différent si les autres attributs ne sont pas pris en compte. Par exemple, pour un aqueduc de 200 de diamètre, un égout sanitaire de 250 de diamètre et un égout pluvial de 450 de diamètre, on peut avoir un taux de productivité réel de 15 m / jour pour un projet, puis de 30 m / jour pour un autre projet si la nature du sol est très différente.

Si l'on résout ce problème en ajoutant une variable d'entrée de la nature du sol et ensuite celui de la météo, de la pelle et ainsi de suite, on se retrouve rapidement avec trop de variables d'entrée et les données sont trop différentes pour permettre un bon apprentissage. Même avec les priorités établies telles qu'illustrées au Tableau 4.2, on remarque après quelques essais (avec les données de démonstration de Weka) qu'au moins six, sept attributs dans l'historique de données touchant à toutes les facettes sont nécessaires. Il faut au moins un attribut qui parle du sol et de l'environnement ainsi qu'un autre qui définit chaque conduite et le couvert par rapport au sol naturel.

Après réflexion, la productivité de la pelle est utilisée pour prendre compte de tout ce qui est environnement et outils (grosueur de pelle, portée de la flèche, nature du sol et élévation du roc...). Cette donnée est idéale, car elle représente bien tous ces attributs en un seul, mais elle n'est pas disponible en avant-plan, seulement lorsque le projet est terminé. Il serait alors nécessaire d'élaborer un deuxième apprentissage indépendant qui pourrait estimer la productivité finale de la pelle en connaissant d'avance les caractéristiques du lieu, de l'environnement et des outils utilisés.

Malgré le peu de données disponibles, il est important de noter que les tests ne sont tout de même pas exécutés sur les mêmes données que l'apprentissage pour éviter de fausser les résultats de l'estimation.

La classification des données est peu homogène. Avec le jeu de plus ou moins 400 journées de données composées d'une dizaine d'attributs, seulement quelques-unes d'entre elles se ressemblent, ce qui a causé un manque d'homogénéité. Autrement dit, on essaie d'estimer une donnée sur laquelle on a appris avec des données qui ont peu de similitudes.

Pour essayer de contrer le problème, il a été nécessaire de reclasser manuellement les données pour bien choisir celles qui utilisées, d'une part, pour l'apprentissage, d'autre part, pour les tests. But principal : éviter une mauvaise séparation des données (si on utilisait le hasard) afin de ne pas se retrouver avec la situation suivante : les trois premiers attributs sont importants tels que définis au Tableau 4.2, soit les diamètres des conduites. Afin de visualiser ce que pourrait engendrer une mauvaise séparation des données si on utilisait le hasard, prenons l'exemple suivant :

Historique de données :

0,300,600,0,4.532,3.010,63.8,11
 0,300,600,0,4.272,2.749,51.07,24
 0,300,600,0,4.28,2.75,50.25,24
 0,300,600,0,4.36,2.84,52.73,12

Historique de données (suite):

0,300,600,0,4.2,2.67,53.84,26
 0,300,600,0,4.09,3.47,56,24
 0,300,600,0,3.91,3.4,61.27,13
 0,300,600,0,4.1,4.02,62.4,24
 0,300,600,0,4.01,4.06,50.72,20
 0,300,900,0,5.98,4.33,33.28,8
 0,300,900,0,5.9,4.38,59.11,15
 0,300,900,0,5.57,4.05,30.93,5
 0,300,900,0,5.39,3.86,16.28,10
 0,300,900,0,4.86,3.34,30.63,11
 0,300,900,0,4.66,3.14,65.02,25
 0,300,900,0,4.455,2.93,53.33,11

Données d'apprentissage :

0,300,600,0,4.532,3.010,63.8,11
 0,300,600,0,4.272,2.749,51.07,24
 0,300,600,0,4.28,2.75,50.25,24
 0,300,600,0,4.36,2.84,52.73,12
 0,300,600,0,4.2,2.67,53.84,26
 0,300,600,0,4.09,3.47,56,24
 0,300,600,0,3.91,3.4,61.27,13
 0,300,600,0,4.1,4.02,62.4,24
 0,300,600,0,4.01,4.06,50.72,20

Données de test :

0,300,900,0,5.98,4.33,33.28,8
 0,300,900,0,5.9,4.38,59.11,15
 0,300,900,0,5.57,4.05,30.93,5
 0,300,900,0,5.39,3.86,16.28,10
 0,300,900,0,4.86,3.34,30.63,11

Données de test (suite):

0,300,900,0,4.66,3.14,65.02,25

0,300,900,0,4.455,2.93,53.33,11

Si la séparation au hasard des données avait généré un tel résultat, on se retrouverait avec une impossibilité d'estimer, puisque l'apprentissage se serait fait sur des données similaires entre elles, mais très différentes des données du test. On remarque dans notre exemple que le premier diamètre représente une conduite d'aqueduc. La valeur 0 veut dire qu'il n'y a pas de conduite d'aqueduc. Le deuxième diamètre représente la conduite d'égout sanitaire. L'historique de données et les données d'apprentissage sont composés de conduites d'égout sanitaire de 300 mm de diamètre. Tandis que la troisième conduite est une conduite d'égout pluvial de 600 mm de diamètre pour l'historique de données et les données d'apprentissage, puis de 900 mm de diamètre pour les données-tests.

Pour résoudre le problème, les données ont été séparées à la main et toutes celles qui n'avaient pas au moins trois autres données similaires (pour l'apprentissage) ont été enlevées. Quel serait le résultat?

Historique de données :

0,300,600,0,4.532,3.010,63.8,11

0,300,600,0,4.272,2.749,51.07,24

0,300,600,0,4.28,2.75,50.25,24

0,300,600,0,4.36,2.84,52.73,12

0,300,600,0,4.2,2.67,53.84,26

0,300,600,0,4.09,3.47,56,24

0,300,600,0,3.91,3.4,61.27,13

0,300,600,0,4.1,4.02,62.4,24

0,300,600,0,4.01,4.06,50.72,20

0,300,900,0,5.98,4.33,33.28,8

0,300,900,0,5.9,4.38,59.11,15

0,300,900,0,5.57,4.05,30.93,5

Historique de données (suite):

0,300,900,0,5.39,3.86,16.28,10
 0,300,900,0,4.86,3.34,30.63,11
 0,300,900,0,4.66,3.14,65.02,25
 0,300,900,0,4.455,2.93,53.33,11

Données d'apprentissage :

0,300,600,0,4.532,3.010,63.8,11
 0,300,600,0,4.272,2.749,51.07,24
 0,300,600,0,4.28,2.75,50.25,24
 0,300,600,0,4.36,2.84,52.73,12
 0,300,600,0,4.2,2.67,53.84,26
 0,300,600,0,4.09,3.47,56,24
 0,300,600,0,3.91,3.4,61.27,13
 0,300,900,0,5.98,4.33,33.28,8
 0,300,900,0,5.9,4.38,59.11,15
 0,300,900,0,5.57,4.05,30.93,5
 0,300,900,0,5.39,3.86,16.28,10
 0,300,900,0,4.86,3.34,30.63,11

Données de test :

0,300,600,0,4.1,4.02,62.4,24
 0,300,600,0,4.01,4.06,50.72,20
 0,300,900,0,4.66,3.14,65.02,25
 0,300,900,0,4.455,2.93,53.33,11

Il est à remarquer que la distribution permet d'apprendre un peu sur chaque groupe avant d'estimer sur deux données du groupe. Le réseau de neurones utilise la rétropropagation pour l'apprentissage. Le réseau se construit automatiquement selon les données de l'historique de données et lorsque, sur la donnée estimée, l'erreur grandit, le réseau s'arrête en divisant la

vitesse d'apprentissage selon le temps restant pour éviter de diverger et assurer la performance du temps d'apprentissage.

Les paramètres de configurations du RNA sont les suivants :

- H = Couches cachées : i = tous les attributs sont des neurones de la couche cachée.
- L = Vitesse d'apprentissage.
- M = Vitesse appliquée au poids pendant la mise à jour.
- N = Temps d'apprentissage (en cycle de branchement et non en seconde).

La fonction d'activation est : $\sigma(v_i) = \tanh(v_i)$ qui est non linéaire limitée par $[-1,1]$.

L'erreur de la rétropropagation est dépendante des données et est calculée en fonction de ceci :

$$e_j(n) = d_j(n) - y_j(n),$$

où j est le nœud de sorti à la donnée n et d est la cible alors qu' y est la valeur produite par le réseau. Le poids du nœud est corrigé en fonction de cette erreur jusqu'à un intervalle accepté (par défaut 10 %).

La règle du delta est déterminée par la règle de base des réseaux de neurones, soit :

$$\Delta w_{ji}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial v_j(n)} y_i(n)$$

Ici, W est le poids (non la donnée d'entrée), Y est la sortie et η , la vitesse d'apprentissage.

Le choix des valeurs des paramètres H, L, M et N du perceptron sont H égale à une couche cachée, qui rejoint les conclusions de Hergazy (1996). Le nombre de neurones dans cette couche est par défaut dans Weka égal au nombre de données d'entrée dans le RNA, ce qui est

également en lien avec Hergazy (1996) et (Moselhi et Siqueira, 1998). L et M sont définis par chacun des essais et N est le pourcentage d'erreur acceptable. Si ce pourcentage n'est pas atteint, cette valeur limite l'apprentissage pour qu'il ne soit pas infini. Elle est par défaut dans Weka à 10 %.

Il est à noter que pour peaufiner un paramètre à sa valeur optimale, il est nécessaire de faire de l'essai / erreur

Voici le Tableau 4.3 présentant les résultats primaires du premier essai. Cet essai avait 361 données d'apprentissage et 27 données-tests.

Tableau 4.3
Résultats primaires du premier essai du réseau de neurones

Attributs	Algorithme	Paramètres	Temps / seconde	Résultats		
				Coefficient de corrélation	Erreur absolue moyenne	Erreur moyenne quadratique
Aqueduc PVC Égout sanitaire PVC Égout pluvial TBA Couvert moy. aqueduc Couvert moy. égout sanitaire Couvert égout pluvial Productivité pelle	Réseau de neurones	-L 0.3 -M 0.2 -N 500 -H i	1,84	0.7533	13.7192	15.4811
		-L 0.03 -M 0.02 -N 1000 -H i	3,2	0.7196	7.7324	9.519
		-L 0.0030 -M 0.0020 -N 10000 -H i	30,74	0.6558	7.615	9.9794
		-L 3.0E-4 -M 2.0E-4 -N 100000 -H i	320,21	0.6585	7.5898	9.9287
		-L 3.0E-5 -M 2.0E-5 -N 1000000 -H i	3230,04	0.659	7.5787	9.9169
	Régression linéaire	Aucun	0,05	0.6372	8.1235	10.2861

Ces résultats ont été utilisés pour déterminer les paramètres optimaux du réseau de neurones. La solution 2, soit -L 0.03 -M 0.02 -N 1000 -H i, apparaît être le meilleur rapport de précision en fonction du temps d'apprentissage, lequel doit être relativement rapide si l'on veut éventuellement introduire cette méthode dans une application en temps réel (ou presque).

Le paramètre de sortie intéressant est l'erreur moyenne absolue (*mean absolute error*), car il nous donne un indice de la différence entre la réalité et la prédiction moyenne du réseau (7.73 veut alors dire qu'en moyenne la réponse se situe à 7.73 m / jour de la vraie réponse).

La différence entre le premier résultat et le deuxième signifie qu'en premier lieu, le réseau n'a pas eu suffisamment de temps pour un apprentissage adéquat. Il atteignait la limite de temps avant la limite de précision. Par contre, pour les résultats subséquents, il devient clair que la précision approche une asymptote à environs 7.5 et que le temps d'apprentissage qui augmente exponentiellement ne vaut pas le mince gain de précision.

Le fait que le temps augmente exponentiellement signifie que le réseau n'atteint jamais l'erreur acceptable de 10 % et ne sort de l'apprentissage que par le temps limite. On aperçoit aussi que nos paramètres nominaux ne donnent pas avantage au réseau de neurones sur la régression linéaire. Dès lors, on doit remettre en question notre approche. Peut-être que les résultats seraient meilleurs si nous avions beaucoup plus de données pour l'apprentissage?

Basé sur les conclusions de Bode (2000) et à la suite des résultats obtenus lors du premier essai, un test est effectué sur un jeu de données beaucoup plus volumineux. Par manque de données réelles, ces dernières seront créées aléatoirement par le chiffrier Excel. Deux modèles de données existants seront utilisés pour générer les données aléatoires :

$$200,250,375,1.80,2.50,2.83,67.20 = 28.09$$

$$250,300,500,1.98,2.35,2.42,56.63 = 28.68$$

Les trois premières valeurs ont été conservées et chaque modèle a été recopié 2 000 fois avec des données aléatoires, pour un total de 4 000 données. Les conduites d'aqueduc, d'égout sanitaire et d'égout pluvial ont toujours comme valeur respectivement 200, 250 et 375 mm de diamètres, puis 250, 300 et 500 mm de diamètre. Le couvert moyen de la conduite d'aqueduc est fixé aléatoirement entre un et deux mètres, le couvert moyen de l'égout sanitaire et de l'égout pluvial est fixé aléatoirement entre deux et quatre mètres, la productivité de la pelle

hydraulique est fixée aléatoirement entre 1 et 70 mètres par jour puis, finalement, le taux de productivité est fixé aléatoirement entre 0 et 40 mètres par jour.

Les tests sont effectués en deux groupes de 2 000 données avec les 400 dernières données utilisées pour test. Les résultats obtenus après une régression linéaire simple sont de 9,9921 et 9,9444 d'erreurs moyennes absolues et, pour le réseau de neurones de paramètres $-L\ 3.0E-2$ - $M\ 2.0E-2$ - $N\ 1000$ - $H\ i$, les résultats sont 9,9443 et 10,12 d'erreurs moyennes absolues.

Si l'on compare ces résultats avec ceux du premier essai, on constate que le nombre de données n'était pas la source du problème puisque nous obtenons un résultat encore plus loin de la réalité, car on a une valeur d'erreur moyenne absolue de 7.7324 pour le résultat accepté du premier essai, puis de 9.9443 pour le meilleur résultat du deuxième essai.

En utilisant une classification nominale, peut-être que le réseau pourrait prendre des décisions plus facilement, puisqu'avec une sortie réelle, il doit utiliser une certaine régression pour déduire une valeur en fonction de celles contenues dans le réseau de neurones.

Suivant ces résultats, il serait intéressant de voir l'impact de la classification du paramètre de sortie en classes nominales. Ce dernier a donc été séparé en classes nominales, soit de a à f séparés comme suit (en mètre) :

a = [0,10[

b = [10,20[

c = [20,25[

d = [25,30[

e = [30,40[

f = [40,60[

Chaque productivité de tranchée a donc été échangée par sa classe respective et les mêmes données ont été utilisées avec les mêmes paramètres afin de comparer les résultats. Lorsque

le réseau de neurones estime des classes, il retourne une matrice de classification plutôt qu'une erreur moyenne, faute d'avoir une valeur réelle à estimer.

Voici la matrice des résultats obtenue avec le classement du paramètre de sortie :

3 1 0 0 0 1 | a = a

0 2 2 0 0 0 | b = b

0 1 2 0 0 0 | c = c

0 0 2 1 1 0 | d = d

0 0 0 0 2 0 | e = e

0 1 0 0 0 3 | f = f

À partir de cette matrice, on peut calculer que 59 % des données sont bien classées, 32 % sont éloignées d'une seule classe et que 9 % sont très mal classées.

Ces résultats ne sont pas facilement comparables à ceux précédents puisqu'ils ne sont pas représentés sous la même forme, donc il est impossible de comparer les deux erreurs moyennes. Il est possible par contre d'évaluer la viabilité de ces résultats. La productivité, 91 % du temps, est bien classée ou alors elle n'est différente que d'une classe. À première vue, cela semble très prometteur, mais vu l'étendue de nos classes (5 à 10 m / jour), cette valeur est peu utile à un estimateur dans le domaine réel d'utilisation. Si les classes étaient plus petites et que les résultats étaient les mêmes, alors cette solution serait très efficace, mais le peu de données réelles disponible ne permet pas de bien vérifier cette affirmation.

Afin de raffiner l'historique de données ou les données d'entrée du réseau de neurones, des experts du domaine ont été consultés afin d'identifier ce qui pourrait être problématique dans les données de l'historique de données ou bien dans les facteurs retenus qui influencent le taux de productivité d'une équipe d'excavation. Les résultats des trois derniers essais dès lors ont été présentés et ils ont fait l'objet de plusieurs remarques. Selon ces estimateurs, estimer le taux de productivité à partir de rapport journalier n'est pas la meilleure méthode. En général, lorsqu'il effectue une évaluation détaillée, les estimateurs se concentrent sur le temps

de productivité totale du projet plutôt que sur le taux de productivité par tronçon exécuté durant une journée. Estimer un projet par tronçon ne permet pas de faire une moyenne globale pour chaque projet. Exemple : si un tronçon est sous-évalué et un autre surévalué ensemble, l'erreur s'annule sur l'ampleur du projet. L'estimation devrait être faite sur le nombre d'heures qu'un tronçon a nécessité pour être effectuée plutôt que de morceler un grand tronçon en journée de travail. Si le projet est la réalisation d'infrastructure dans des rues existantes, le temps de travail sera plus long. La pente d'excavation est plus importante que la nature du sol puisqu'elle déterminera l'action de la pelle; une pente raide nécessite peu de travail de la pelle, alors qu'une pente faible nécessite beaucoup plus de travail de la pelle. La pente est déterminée par la nature du sol. La météo n'affecte pas le travail à moins de conditions extrêmes et, à ce moment, les travaux seront suspendus.

Suivant les discussions avec les experts, certains changements ont été effectués aux données d'entrée. Pour ce qui est des recommandations qui concernent les données de l'historique de données, elles pourront servir pour l'amélioration des rapports journaliers des surveillants de chantier afin d'obtenir les données sous une forme acceptable pour les réseaux de neurones. Un paramètre « existant » a été ajouté à chaque tranchée. Ce paramètre a deux valeurs possibles : oui ou non. Le paramètre « nature du sol » a été remplacé par la pente : faible, moyenne ou forte. La durée de travail par tranchée est maintenant la donnée à estimer. Les données de tests seront tous les tronçons de tranchée d'un projet existant et d'un projet non existant. Ces deux projets ont été choisis en fonction de leur similarité avec les projets d'apprentissage pour éviter des tests sur des projets uniques. Le paramètre de sortie est un réel, car il est inutile de le séparer en classes comme précédemment.

Voici les résultats obtenus avec le réseau de neurones et la nouvelle méthode :

<i>Coefficient de corrélation</i>	-0.1346
<i>Erreur absolue moyenne</i>	1.7851
<i>Erreur relative quadratique</i>	2.1057
<i>Erreur absolue relative</i>	99.8772 %
Erreur relative quadratique	99.9632 %
<i>Nombre total de tronçons de tranchées</i>	42.0

Avec la nouvelle méthode et la régression (pour comparaison) :

<i>Coefficient de corrélation</i>	-0.1605
<i>Erreur absolue moyenne</i>	1.8576
<i>Erreur moyenne quadratique</i>	2.2458
<i>Erreur absolue relative</i>	103.9313 %
<i>Erreur relative quadratique</i>	106.6140 %
<i>Nombre total de tronçons de tranchées</i>	42.0

Ces derniers résultats sont nettement supérieurs aux meilleurs résultats obtenus jusqu'à présent, soit 1,7851 d'erreur moyenne (en heures) par tranchée d'environ 10 h, comparés à 7.732 m / jour par tranchée d'environ 40 m / jour.

De plus, cette solution dépasse même la régression, ce qui prouve que notre réseau tire maintenant avantage des données d'entrée nominales telles la pente d'excavation et l'existence du projet.

4.4 Application du modèle à l'ensemble des données avec les données d'entrée révisées

Afin de réduire davantage le nombre de paramètres d'entrée, des changements ont été appliqués. La distance entre les différentes conduites a été retirée, car elles ne varient pratiquement que lors de la réalisation des travaux. Le couvert moyen de la conduite

d'aqueduc a également été retiré, car la variation de ce paramètre est marginale. Les paramètres d'entrée des conduites ont été restructurés en deux valeurs d'entrée. Le premier étant le nombre de conduites et le second, le diamètre de la conduite d'égout pluvial sous la forme suivante :

0	→ aucun
1	→ <600
2	→ [600-900[
3	→ [900-1200[
4	→ >1200

En ce qui concerne les différents accessoires installés dans la tranchée lors des travaux, ils ont été restructurés. Tous les éléments sont représentés en une seule valeur d'entrée de complexité. Cette complexité est calculée par le total d'heures attribuées à chaque élément. Voir la description suivante :

1 vanne	= 0 h
1 chambre de vanne	= 1 h
1 regard	= 1 h
1 entrée de service	= 1,5 h
1 puisard	= 1,5 h
1 borne-fontaine	= 2 h
1 raccord	= 2,5 h

Et le total des heures a été traduit en complexité de la tranchée selon la distribution suivante :

0	= aucune
]0,4[= simple
[4,8[= moyenne
[8,12[= difficile
[12,∞[= complexe

Les résultats obtenus avec les valeurs d'entrée redéfinies sont pour le réseau de neurones.

<i>Coefficient de corrélation</i>	0.2504
<i>Erreur absolue moyenne</i>	1.6446

Et les résultats obtenus avec les valeurs d'entrée redéfinies sont pour la régression linéaire les suivants :

<i>Coefficient de corrélation</i>	-0.0392
<i>Erreur absolue moyenne</i>	1.9324

Les résultats obtenus avec les deux projets mesurés avec les nouveaux paramètres sont acceptables et ont été distribués à trois estimateurs experts afin de comparer leurs résultats à celui du réseau de neurones.

4.5 Interprétation des résultats

La comparaison des résultats se fait d'un projet à l'autre afin d'être représentative de la réalité. Il est à noter qu'ils ne représentent pas l'erreur par tranchée, mais bien l'erreur par projet. Autrement dit, si une tranchée d'un projet est surestimée de 50 heures et qu'une autre est sous-estimée de 50 heures aussi, alors leur erreur s'annulera au niveau du projet. Les résultats comparatifs sont présentés au Tableau 4.4, lequel est divisé en deux parties afin de comparer les heures obtenues avec les heures réelles et de comparer le pourcentage d'erreur obtenue par rapport aux résultats réels émanant des rapports de chantier des surveillants de travaux.

Tableau 4.4

Tableau comparatif des résultats finaux en heures et en pourcentage d'erreur par projet

Tableau comparatif des résultats
des durées en heures par projet

Projet	Réelle	RNA	Régression	Humain 1	Humain 2	Humain 3
720-134	363 h	351,277	305,446	231,37	288,6	252
788-134	69.5 h	71,444	63,499	83,77	72,4	90

Tableau comparatif des pourcentages d'erreur
par projet par rapport à la durée réelle

Projet	RNA	Régression	Humain 1	Humain 2	Humain 3
720-134	-3,2 %	-15,9 %	-36,3 %	-20,5 %	-30,6 %
788-134	2,8 %	-8,6 %	20,5 %	4,2 %	29,5 %

Ci-après, les résultats par tronçons de tranchée sont présentés au Tableau 4.5, lequel représente les différences obtenues pour chacun des tronçons avec les RNA et à partir des mesures des estimateurs chevrons. La première colonne représente les différents tronçons; la deuxième, le taux de productivité réel obtenu à partir des rapports journaliers des surveillants de chantier; la troisième, le taux de productivité estimé par le RNA; finalement, les deux dernières colonnes, les différences en heures puis en pourcentage. Le premier groupe de résultats concerne le projet dans une rue existante, tandis que le groupe du bas, celui dans un terrain vacant.

Tableau 4.5

Tableau comparatif des résultats par tronçons de tranchée pour le réseau de neurones

« Durée »

Projet 720-134

Tranchée	Durée réelle (h)	Estimation neurones	Estimation linéaire (h)	Estimation humain 1 (h)	Estimation humain 2 (h)
TR720001	10.5	5.68	7.84	5.13	4.00
TR720002	10.5	11.29	9.02	9.39	14.50
TR720003	10.5	10.43	9.13	8.54	12.50
TR720004	6.5	3.32	7.65	1.28	2.50
TR720005	10.5	3.26	7.96	4.17	5.00
TR720006	5	10.26	8.63	4.15	5.00
TR720007	10.5	10.44	8.89	4.49	7.00
TR720008	10.5	5.69	7.92	2.54	4.50
TR720009	10.5	10.35	9.31	7.22	12.50
TR720010	10.5	10.63	8.79	6.35	7.50
TR720011	12	10.73	8.60	3.21	3.90
TR720012	11	10.36	9.29	9.14	12.00
TR720013	10.5	12.66	9.18	9.77	11.40
TR720014	10.5	10.41	9.51	11.38	13.00
TR720015	12	12.10	8.65	7.50	9.50
TR720016	10.5	10.79	8.38	9.62	10.00
TR720017	14	11.12	8.47	3.17	3.00
TR720018	10	10.67	8.68	3.76	4.50
TR720019	12.5	10.36	9.27	8.51	12.00
TR720020	11.5	10.51	7.91	5.51	5.40
TR720021	11.5	11.18	8.07	6.34	6.70
TR720022	13.5	12.74	8.68	7.99	10.90
TR720023	11	10.61	9.21	9.37	11.70
TR720024	11	11.07	8.13	6.03	8.00
TR720025	10	10.82	8.00	4.76	5.70
TR720026	11	10.63	9.60	7.59	11.00
TR720027	10	10.35	9.00	8.18	9.00
TR720028	5.5	10.18	8.82	6.72	6.80
TR720029	10	10.17	9.01	8.39	9.00
TR720030	12.5	10.75	7.96	4.86	5.00
TR720031	9.5	10.18	9.01	7.57	7.30
TR720032	10	10.61	9.39	10.69	8.50
TR720033	8.5	10.31	9.20	5.17	9.80
TR720034	9.5	10.32	9.14	6.30	10.00
TR720035	9.5	10.32	9.16	6.58	9.50
TOTAL	363	351.28	305.45	231.37	288.60

Tableau 4.5 (suite)

Tableau comparatif des résultats par tronçons de tranchée pour le réseau de neurones

« Différence entre la durée réelle et les résultats en pourcentage »

Tranchée	%Δ neurones	%Δ linéaire	%Δ humain 1	%Δ humain 2
TR720001	-45.9%	-25.3%	-51.1%	-61.9%
TR720002	7.6%	-14.1%	-10.6%	38.1%
TR720003	-0.7%	-13.1%	-18.7%	19.0%
TR720004	-49.0%	17.7%	-80.3%	-61.5%
TR720005	-69.0%	-24.2%	-60.3%	-52.4%
TR720006	105.1%	72.6%	-17.0%	0.0%
TR720007	-0.5%	-15.3%	-57.2%	-33.3%
TR720008	-45.8%	-24.6%	-75.8%	-57.1%
TR720009	-1.4%	-11.4%	-31.2%	19.0%
TR720010	1.2%	-16.3%	-39.5%	-28.6%
TR720011	-10.6%	-28.3%	-73.3%	-67.5%
TR720012	-5.8%	-15.5%	-16.9%	9.1%
TR720013	20.5%	-12.6%	-7.0%	8.6%
TR720014	-0.8%	-9.4%	8.4%	23.8%
TR720015	0.8%	-27.9%	-37.5%	-20.8%
TR720016	2.8%	-20.2%	-8.4%	-4.8%
TR720017	-20.6%	-39.5%	-77.4%	-78.6%
TR720018	6.6%	-13.3%	-62.4%	-55.0%
TR720019	-17.1%	-25.8%	-31.9%	-4.0%
TR720020	-8.6%	-31.2%	-52.1%	-53.0%
TR720021	-2.8%	-29.9%	-44.9%	-41.7%
TR720022	-5.7%	-35.7%	-40.8%	-19.3%
TR720023	-3.6%	-16.3%	-14.8%	6.4%
TR720024	0.6%	-26.1%	-45.2%	-27.3%
TR720025	8.1%	-20.0%	-52.4%	-43.0%
TR720026	-3.4%	-12.7%	-31.0%	0.0%
TR720027	3.5%	-10.0%	-18.2%	-10.0%
TR720028	85.0%	60.4%	22.2%	23.6%
TR720029	1.6%	-9.9%	-16.1%	-10.0%
TR720030	-14.0%	-36.4%	-61.1%	-60.0%
TR720031	7.1%	-5.1%	-20.3%	-23.2%
TR720032	6.1%	-6.1%	6.9%	-15.0%
TR720033	21.3%	8.3%	-39.2%	15.3%
TR720034	8.7%	-3.7%	-33.7%	5.3%
TR720035	8.7%	-3.5%	-30.7%	0.0%
TOTAL	-3.2%	-15.9%	-36.3%	-20.5%

Tableau 4.5 (suite)

Tableau comparatif des résultats par tronçons de tranchée pour le réseau de neurones

« Durée »

Projet 788-134

Tranchée	Durée réelle (h)	Estimation neurones	Estimation linéaire (h)	Estimation humain 1 (h)	Estimation humain 2 (h)
TR788001	10	12.44	9.02	11.52	9.60
TR788002	9.5	12.93	9.08	12.72	11.20
TR788003	10	12.84	9.42	18.24	14.60
TR788004	10	2.51	8.50	4.35	4.00
TR788005	10	10.13	9.15	10.07	7.90
TR788006	10	10.25	9.03	10.75	10.10
TR788007	10	10.35	9.30	16.12	15.00
TOTAL	69.5	71.44	63.50	83.77	72.40

« Différence entre la durée réelle et les résultats en pourcentage »

Tranchée	%Δ neurones	%Δ linéaire	%Δ humain 1	%Δ humain 2
TR788001	24.4%	-9.8%	15.2%	-4.0%
TR788002	36.1%	-4.5%	33.9%	17.9%
TR788003	28.4%	-5.8%	82.4%	46.0%
TR788004	-75.0%	-15.0%	-56.5%	-60.0%
TR788005	1.3%	-8.5%	0.7%	-21.0%
TR788006	2.5%	-9.7%	7.5%	1.0%
TR788007	3.5%	-7.0%	61.2%	50.0%
TOTAL	2.8%	-8.6%	20.5%	4.2%

Avec les résultats obtenus précédemment, il semble que le réseau de neurones surpasse les prédictions des humains ainsi que de la régression linéaire. Il est possible d'affirmer que le fait que les tranchées hautement erronées s'annulent entre elles aide le réseau de neurones à s'approcher de la réalité. En fait, ce phénomène néanmoins s'applique à tous les autres aussi. Deux des estimateurs humains ont envoyé leurs résultats par tranchée et ce phénomène est aussi visible dans leurs résultats.

Si l'on mesurait l'erreur absolue des tranchées par projet, on obtiendrait tout de même un meilleur résultat pour le réseau de neurones comme l'indique l'erreur absolue moyenne donnée par Weka : 1,6446.

Si on compare cette erreur absolue moyenne à la moyenne du temps des tranchées (10,3 heures), on obtient 16 % de marge d'erreur. En appliquant cette marge d'erreur aux totaux des projets, on obtient les marges suivantes :

Projet de 363 heures : [304,92 à 421,08]

Projet de 69,5 heures : [58,38 à 80,62]

Revenons sur les résultats. Malgré cette marge, le réseau de neurones demeure le grand champion puisqu'aucun humain n'atteint la marge à l'exception de l'humain 2 pour le projet de 69,5 heures. En outre, après vérification, cette personne avait réalisé les travaux de ce projet dans le passé, donc le résultat est erroné. La régression se situe dans la marge pour les deux projets, mais en bordure de celle-ci.

En conclusion, la seule question probable pour mettre en doute le résultat est celle-ci : est-ce que les résultats seraient semblables à d'autres projets? La réponse dépend du nombre de données disponibles pour l'apprentissage, mais si la plupart des tranchées sont similaires à d'autres, déjà étudiées par le réseau, alors le résultat sera prometteur.

CONCLUSION

Le but de cette recherche était de répondre à la question « Est-ce qu'un modèle de RNA éprouvé a une meilleure qualité de prédiction que des estimateurs chevronnés pour l'attribution du taux de productivité d'une équipe de travail lors de l'utilisation de la méthode d'estimation? »

Pour répondre à cette question, premièrement on a défini les différentes caractéristiques et limites de la méthode d'estimation détaillée à partir des spécifications de l'AACE. Cet organisme a été choisi parmi plusieurs autres qui définissent différentes spécifications pour la méthode d'estimation détaillée, car cette association a une vision plus spécifique du domaine de la construction et un rayonnement plus international. On a vu que selon l'AACE la précision de cette méthode d'estimation est de 30 % @ -15 %. Cette fluctuation est majeure si l'on prend en considération que le pourcentage de profit d'une entreprise est de l'ordre de 5 %.

Il a également été démontré que l'attribution du taux de productivité d'une équipe de travail est la source majeure d'erreur pour ce type d'estimation.

À partir des rapports journaliers des surveillants de chantier de 41 projets exécutés par une firme de génie-conseil de la région de Laval, on a créé un historique de données. Cet historique de données contenait les valeurs influençant le taux de productivité d'une équipe de travail ainsi que les taux de productivité obtenus sur le chantier pour chaque journée de travail des 41 projets. Les facteurs influençant le taux de productivité ont été déterminés en référence à des discussions avec 3 estimateurs expérimentés du domaine.

Les données brutes émanant des rapports journaliers demeurent une possibilité d'erreur, car ils dépendent du niveau de connaissance et du niveau de rigueur du surveillant et du supérieur immédiat de la firme de génie.

Avant de pouvoir effectuer des essais sur les RNA, il fallait déterminer le jeu des données d'entrées au réseau de neurones. Plus de 4 essais ont été nécessaires pour en arriver à des résultats satisfaisants. Selon les experts du domaine, il existait plus d'une quinzaine de facteurs influençant le taux de productivité. Après analyse, certains de ces facteurs en cachaient d'autres et finalement on avait plus d'une vingtaine de données d'entrée. Ce nombre rendait difficile la paramétrisation du RNA. Après plusieurs essais et en se basant sur les conclusions de Setyawati et Creese (Setyawati, Sahirman et Creese, 2002), on a diminué ce nombre à moins de 10 avec lesquels on a obtenu des résultats satisfaisants.

Après l'entraînement du RNA, les valeurs d'entrée des 2 projets restants ont été utilisées afin de prédire le taux de productivité de ces projets dans le but des comparés aux taux réels obtenus au chantier. On a également demandé à 3 estimateurs du domaine de calculer à partir des plans et devis de ce projet quel serait selon eux le taux de productivité de l'équipe de tranchées de ces 3 projets. Les résultats obtenus sont présentés au Tableau no 5.1 ici-bas.

Tableau 5.1

Tableau comparatif des résultats finaux en heures et en pourcentage d'erreur par projet

Tableau comparatif des résultats
des durées en heures par projet

Projet	Réelle	RNA	Régression	Humain 1	Humain 2	Humain 3
720-134	363 h	351,277	305,446	231,37	288,6	252
788-134	69.5 h	71,444	63,499	83,77	72,4	90

Tableau comparatif des pourcentages d'erreur
par projet par rapport à la durée réelle

Projet	RNA	Régression	Humain 1	Humain 2	Humain 3
720-134	-3,2 %	-15,9 %	-36,3 %	-20,5 %	-30,6 %
788-134	2,8 %	-8,6 %	20,5 %	4,2 %	29,5 %

Ces résultats sont très prometteurs, car le taux de productivité global des projets se situe entre 2.8 % @ -3.2 % pour les réseaux de neurones ce qui est bien à l'intérieur des spécifications de la norme de AACE no 17R-97 qui spécifie qu'une estimation de classe 1 et 2 doit avoir une précision se situant entre 30 % @ -15 %. Ces résultats ne sont pas seulement à l'intérieur des spécifications, mais les améliorent de façon nette.

Si l'on considère que le taux de productivité influence la valeur totale du coût des ressources d'équipements et de main d'œuvre et que ces ressources comptent pour environ 50 % du coût total du projet, on peut considérer que ces résultats annoncent un plus pour l'estimation détaillée des coûts de projet.

En outre, si l'on compare les résultats des RNA avec les résultats des estimateurs chevronnés on peut considérer que les résultats des RNA sont supérieurs en précision de façon claire. La seule valeur obtenue par un estimateur qui est près des résultats des RNA est le 4.2 % obtenus par l'estimateur no 2. On rappelle qu'après vérification auprès de cet estimateur, celui-ci a participé à la construction de ce projet en tant qu'estimateur et gérant de projet. Ce résultat ne peut donc pas être tenu en compte.

Les données sources ont été transformées afin de fournir des données propres à l'utilisation avec des RNA. Il va de soi que les rapports journaliers de chantier pourraient être modifiés afin de fournir les données du chantier sous une forme acceptable pour les valeurs d'entrée du RNA. On pourrait également mettre en place un processus de vérification des données compilées dans ces rapports.

On rappelle qu'il a été difficile de traiter les données sources, car les données représentaient le travail d'une journée seulement. Si l'équipe de travail a effectué plusieurs types de travaux, le taux de productivité pour chacun de ces types de travaux est noté que pour de petites quantités réalisées. Ce fait peut faire en sorte de fausser le taux de productivité de l'ensemble d'un tronçon qui aurait pris plus de 7 @ 8 jours de travail. Des règles d'affaires pourraient être créées afin de compiler à partir des rapports journaliers, les résultats par tronçon spécifique. Dans ces rapports, le taux de productivité de chacun des accessoires pourraient

être notés afin d'éliminer toutes ambiguïtés et interprétation nécessaire avec des rapports journaliers utilisés. On rappelle que l'importance de la valeur du taux de productivité, lors de l'estimation détaillée, est d'obtenir un taux le plus précis possible pour l'ensemble du projet et pour chacune des activités à réaliser.

On se rend compte que les RNA peuvent être utilisés pour la prédiction des taux de productivité des équipes de travail. La majorité du travail se situe dans la détermination des valeurs d'entrées du RNA qui influencent ce taux de productivité.

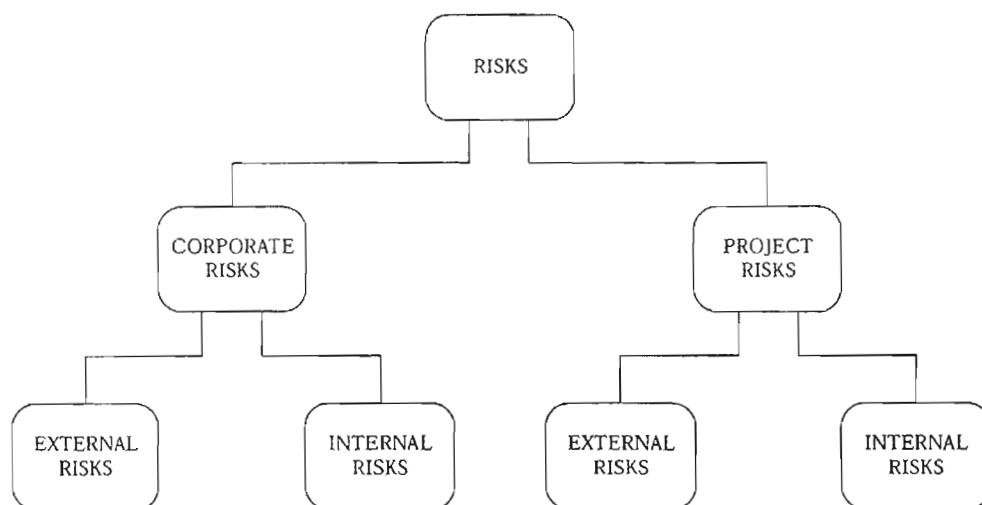
Des recherches futures pourraient être effectuées pour déterminer si la précision peut être influencée par un historique de données plus volumineux. De plus, il serait intéressant d'explorer si un réseau de neurones peut être entraîné pour représenter les performances d'une entreprise. Pour ce faire, il faudrait utiliser plusieurs projets effectués par une entreprise A puis plusieurs autres projets exécutés par une entreprise B. Le RNA1 pourra être entraîné avec les projets de l'entreprise A. Et le RNA2 pourra être entraîné avec les projets de l'entreprise B. Par la suite, il serait intéressant de vérifier quelle précision on obtient avec des projets de l'entreprise A et le RNA 2 et vice versa. Ceci permettrait de vérifier si le RNA peut être spécifique à une entreprise ce qui serait très apprécié par les propriétaires d'entreprise soumis au jeu des soumissions publiques.

Un des objectifs était de vérifier les performances informatiques afin de vérifier la possibilité d'introduire des systèmes de prédictions dans les logiciels commerciaux d'estimation détaillée du coût de projet. Le coût temps nécessaire à la prédiction se mesurait en quelques secondes ce qui est très acceptable. La majorité du temps passé sur les RNA se concentre sur la paramétrisation et l'entraînement du RNA.

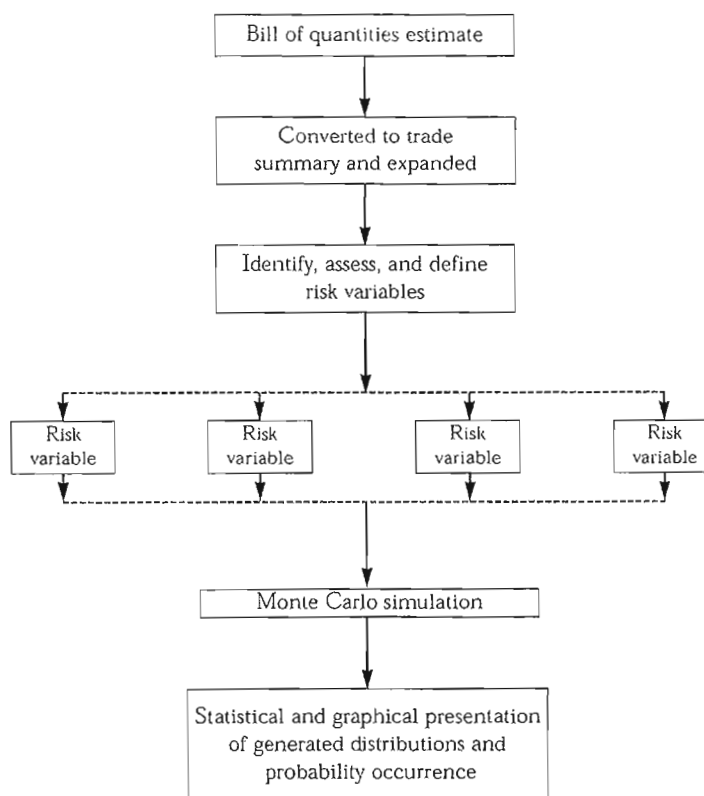
APPENDICE A

INFORMATIONS COMPLÉMENTAIRES DE LA REVUE DE LITTÉRATURE

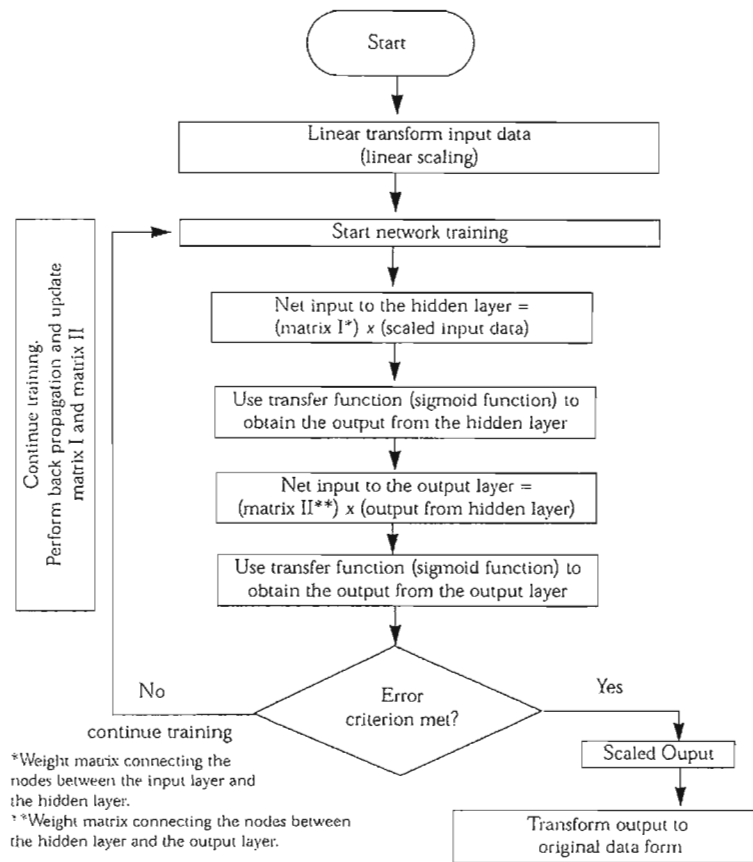
A.1	The General Risk Classification Method	78
A.2	The Proposed Probabilistic Cost Estimating Model.....	78
A.3	Neural Network Training for Development of Weight Matrices I and II	79
A.4	General Structure of Neural Network in Timber Bridge Cost Estimation	79



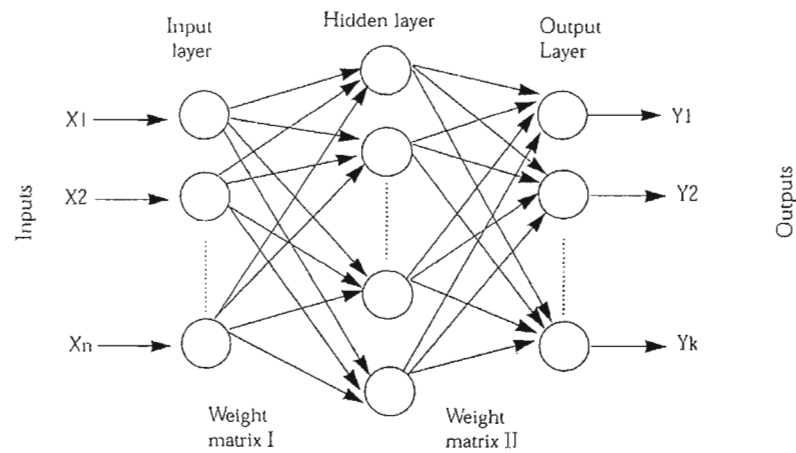
Appendix A.1 - The General Risk Classification Method



Appendix A.2 - The Proposed Probabilistic Cost Estimating Model



Appendice A.3 - Neural Network Training for Development of Weight Matrices I and II



Appendice A.4 - Gneral Structure of Neural Network in Timber Bridge Cost Estimation

BIBLIOGRAPHIE

- Adeli, Hojjat. 2001. «Neural Networks in Civil Engineering: 1989-2000». *Computer-Aided Civil and Infrastructure Engineering*. vol. 16, p. 126-142.
- Akintoye, As; Maclead, M.J. 1997. «Risk and Its Management In Construction». *International Journal of Project Management*. vol. 15, p. 31-38.
- Al-Tabtabai, Hashem, Alex P. Alex et Maha Tantash. 1999. «Preliminary Cost Estimation of Highway Construction Using Neural Networks». *Cost Engineering*. vol. 41, no 3.
- Bode, Jürgen. 1998. «Neural Networks for Cost Estimation». *Cost Engineering*. vol. 40, no 1.
- Bode, Jürgen. 2000. «Neural networks for cost estimation: simulations and pilot application». *International Journal of Production Research*. vol. 39 no 6, p. 1231-1254.
- Cooper, D.R. ; Schindler, P.S. 2002. *Business Reserch Méthods*, Eight Edition. New York.
- Creese, Robert C., et Li Li. 1995. «Cost Estimation of Timber Bridges Using Neural Networks». *Cost Engineering*. vol. 37, no 5.
- Dysert, R. Larry CCC. 2004. «Estimating». In *Skills & Knowledge of Cost engineering*, 5 th Edition, AACE.
- El-Gafy, Mohamed A., Mahmoud A. Taha et Moheeb El-Said. 2000. «Neural Network Model for Parametric Cost Estimation of Sewer Projects».
- Evans, Donald L.; Kassiger, Theodore W.; Cooper, Kathleen B.; Kincannon Charles Louis (2004). Highway, Street, and Bridge Construction: 2002. U.S. Department of Commerce, USCENSUSBUREAU En ligne.
<www.census.gov/const/www/cci/fintro.html>.
- Genest, Bernard-André; Nguyen, Tho, Hau. 2002. *Principes et techniques de la gestion de projets*, Édition 3.
- Hamilton, Allen C. CCE, et Richard E. PE Westney. 2002. «Cost Estimating Best Practices». *AACE International Transactions*. vol. EST.02.
- Han, Jiawei; Kamber, Micheline. 2001. *Data Mining-Concepts and Techniques*.
- Hancher, Dr. Donn E.; Abd-Elkhalek, Dr. Hesham A. 1998. «The Effect of Hot Weather on Construction Labor Productivity and Costs». *Cost Engineering*. vol. 40/no4. Hegazy,

- T., P. Fazio et O. Moselhi. 1994. «Developing Practical Neural Network Applications Using Back-Propagation». *Microcomputers in Civil Engineering*. vol. 9, no 2, p. 145-159.
- Hegazy, Tarek, et Osama PE Moselhi. 1995. «Elements of Cost Estimation: A Survey in Canada and United States». *Cost Engineering*. vol. 37, no 5.
- Kim, Gwang-Hee, Sung-Hoon An et Kyang-In Kang. 2004. «Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning». *Science direct Building and Environment*. vol. 39, p. 1235-1242.
- Kim, Gwang-Hee, Jie-Eon Yoon, Sung-Hoon An et Cho, Hun-Hee
- Kang, Kyung-In. 2004. «Neural network model incorporating a genetic algorithm in estimating construction costs». *Science direct Building and Environment*. vol. 39, p. 1333-1340.
- Leonard, Charles A., Paul Fazio et Osama PE Moselhi. 1988. «Construction Productivity: Major Causes of Impact». *AACE International Transactions*. vol. D-10.
- Leu, Sou-Sen; Lo, Hsien-Chuang. 2004. «Neural-network-based regression model of ground surface settlement induced by deep excavation». *Automation in Construction*. vol. 13, p. 279-289.
- Li, Shuhui, Donald C. Wunsch, Edgar O'Hair et Michael G. Giesselmann. 2001. «Comparative Analysis of Regression and Artificial Neural Network Models for Wind Turbine Power Curve Estimation». *Journal of Solar Energy Engineering*. vol. 123, p. 327-332.
- Lippman, R.P. 1987. «An Introduction To Computing With Neural Nets». *IEEE Acoustics, Speech and Signal Processing Magazine*. vol. 4, p. 4-22.
- Lu, Ming, Daniel S. Yeung et Wing W. Y. Ng. 2005. «Applying undistorted neural network sensitivity analysis in iris plant classification and construction productivity prediction». *Soft Comput.* vol. 10, p. 68-77.
- McCally, Bob M. 1999. «Demonstrated Labor Efficiency: An Effective Cost Control and Analytical Tool». *Cost Engineering*. vol. 41.
- McKim, Robert A. PE. 1993. «Neural Network Applications to Cost Engineering». *Cost Engineering*. vol. 35, no 7.
- Moselhi, Osama PE, Tarek Hegazy et Paul Fazio. 1991. «Practical Cost Estimation in a Competitive Environment». *AACE International Transactions*. vol. H.5.
- Moselhi, Osama PE, et Ines Siqueira. 1998. «Neural Networks for Cost Estimating of Structural Steel Buildings». *AACE International Transactions*. vol. IT/IM.06.

- Piatetsky-Shapiro, G.; Frawley, W.J. 1991. «Knowledge Discovery in Databases». *AAAI/MIT Press*.
- PMBOK, Guide. 2004. «Guide du Corpus des connaissances en management de projet».
- Popham, Kymberli. 1996. «Cost Estimating Using Historical Costs». *AACE International Transactions*, no 14.
- Rowings Jr., James E. PE CCE, et Rifat Sonmez. 1996. «Labor Productivity Modeling With Neural Networks». *AACE International Transactions*. vol. Prod.1.
- Setyawati, Bina R., Robert C. Dr PE CCE Creese et Sidharta Sahirman (2003). Neural Networks for Cost Estimation (Part 2). *AACE International Transactions*. EST.14
- Setyawati, Bina R., Sidharta Sahirman et Robert C. Dr PE CCE Creese. 2002. «Neural Networks for Cost Estimation (Part 1)». *AACE International Transactions*. vol. EST.13.
- Singh, Amarjit. 1993. «Non-linear Earned Value Analysis». *AACE International Transactions*, no C.16.
- Smith, Alice E., et Anthony K. Mason. 1996. «Cost Estimation Predictive Modeling: Regression versus Neural Network». *Journal The Engineering Economist*.
- Uher, Dr Thomas E. 1996. «A Probabilistic Cost Estimating Model». *Cost Engineering*. vol. 38, no 4.
- Yu, Wen-der; Lai, Chien-chung; Lee, Wan-li. 2005. «A WICE approach to real-time construction cost estimation». *Automation in Construction* 15, p. 12-19.